Lectures for MAT 394: Forensic DNA Analysis

Jay Taylor

March 18, 2013

Contents

Chapter 1

Overview and History

1.1 A Brief History of Forensic Genetics

Forensic protein profiling was introduced well before DNA sequencing techniques were developed and uses phenotypic variation as an imperfect proxy for underlying genetic polymorphism. Examples include the ABO blood group system and allozyme variation.

- The ABO blood group system was described by Karl Landsteiner in 1900 and first used as forensic evidence in Italian courts around 1915.
- Four detectable phenotypes: A, B, AB, and O with approximate global frequencies of 32%, 22%, 5%, and 40%, respectively.
- Easy to assay using serology.
- Blood group can only be used to exclude possible suspects.
- Allozymes are charge/length variations in proteins that are detectable using gel electrophoresis. Limited variation again means that these can be used for exclusion, but not identification.

DNA fingerprinting for forensic purposes was proposed by Alec Jeffreys (1985). This saw its first application in 1986 to a double murder/rape case in the English Midlands.

- Two 15-year old girls, Lynda Mann and Dawn Ashworth, were raped and murdered near Narborough, Leicestershire in 1983 and 1986, respectively.
- Police arrested a 17-year-old local suspect, Richard Buckland, who confessed to the second murder but denied involvement in the first.
- DNA profiling using semen samples from both crime scenes excluded Buckland as the culprit.
- Blood samples were volunteered by over 4000 local men, but none were found to match the DNA recovered at the crime scenes.
- In 1987, a man was heard boasting that he had been paid to provide a sample for a local baker, Colin Pitchfork.
- Pitchfork was arrested and testing showed that his DNA was a match to that from both crimes. He subsequently confessed to and was convicted of both crimes.

Some other key developments and milestones include:

- 1988: The FBI begins using RFLP profiles in its casework.
- 1989: DNA evidence was excluded during pretrial arguments in *NY v. Castro* because of concerns about quality assurance in the lab processing the data.
- 1989 - 1995: 'DNA Wars' involve a series of court cases and publications questioning the reliability of DNA evidence. Statistics and population genetics are central to these disputers.
- 1992, 1996: NRC I and II Reports are published laying out guidelines for the processing and interpretation of DNA evidence.
- 1995: UK DNA Database is established by the Forensic Science Services.
- 1998: US Combined DNA Index System (CODIS) is established by the FBI.

1.2 Evidence and Statistics: Some Preliminary Examples

The following examples come from Chapter 2 of Balding (1995) and illustrate some of the statistical considerations that come into play when interpreting DNA evidence.

Example 1.1. ***Positive predictive value of a diagnostic test****. Suppose that a diagnostic test is available for a rare disease with the following properties.*

- *The incidence of the disease is 0.1%, i.e., approximately 1 in 1000 people will get the disease.*
- *The test is highly specific: the false-positive rate is 1%.*
- *The test is sensitive: the fale-negative rate is 5%.*

Assuming that you have a positive diagnosis with this test, what is the probability that you have the disease?

We can reason as follows. Consider a population containing 100,000 people. Of these, approximately 100 will have the disease and of these 100, 95 will have a positive test result and 5 will have a negative test result. Of the remaining 99,900 people without the disease, 999 will also receive a positive test result. Thus, on average, the proportion of true positives among all those who test positive for the disease (the PPV) will be about $95/1094 \approx 0.087$*. In this case, although the test is very accurate, the disease is sufficiently rare that a positive test result is more likely to be a false positive than to be a correct diagnosis.*

Example 1.2. ***Crime on an island.*** *Suppose that a crime is committed on an island with a population of 101 and that the culprit is known to have a rare trait* $\mathcal{T}$*. Assume that:*

- *The culprit is known to be from the island and all islanders are initially under equal suspicion.*
- *A suspect is arrested who has trait* $\mathcal{T}$*, but there is no other evidence against the suspect.*
- *The presence or absence of* $\mathcal{T}$ *in the remaining islanders is unknown.*

- *Based on studies of $\mathcal{T}$ on the mainland, the trait is expected to occur in a fraction $p = 0.01$ of the population and this fraction is not changed by the fact that the suspect has the trait.*

How strong is the evidence that the suspect is the culprit?

As in the first example, the predictive value of the trait can be quantified by reasoning that, on average, $100 \times 0.01 = 1$ other person on the island will have the trait and that the true culprit is equally likely to be either of the two individuals with $\mathcal{T}$, i.e., conditional on the available evidence, the probability that the suspect is the true culprit is $1/(1+1) = 0.5$. Thus, even though the trait is fairly rare, on its own it does not provide 'overwhelming' evidence against the suspect. Later in the course, we will show that if the frequency of the trait is p, then the ***posterior probability*** *that the suspect is guilty is equal to*

$$\mathbb{P}(G|E) = \frac{1}{1+Np}.$$

Here G stands for the proposition that the suspect is guilty and E denotes the available evidence, namely, that both the suspect and the culprit have trait $\mathcal{T}$.

In practice, there are many complicating factors that were ignored in Example 2.1 that have the potential to significantly alter the weight of the evidence against the suspect. Some of these issues were the subject of the debates concerning the appropriate interpretation of DNA evidence that took place in the early 1990's and which were addressed by the two NRC reports.

Uncertainty about trait frequencies: If the frequency of $\mathcal{T}$ is only known imperfectly, e.g., the island population may be genetically differentiated from the mainland population, then it can be shown that

$$\mathbb{P}(G|E) = \frac{1}{1+N(p+\sigma^2/p)},$$

where σ^2 is the variance of the estimate of p. When p is small (as is usually desirable of traits that will be used in identification), σ can be of the same order of magnitude as p itself. Furthermore, because this probability is a decreasing function of σ^2, analyses that underestimate or ignore this source of uncertainty will tend to be prejudiced against the suspect. For example, if $p = 0.01$ and $\sigma = 0.01$, then $\mathbb{P}(G|E) = 1/(1+2) = 0.3333$, which is significantly smaller than the estimate obtained when it is assumed that p is known exactly.

Relatedness: An important complication that arises when assessing DNA evidence is the non-independence of the genotypes of a suspect and their relatives. In particular, the observation that a suspect has a given DNA profile increases the probability that other individuals in the population, especially those that are closely related to the suspect, also have that profile. In the most extreme case that the suspect has a genetic twin, this probability is close to one. This information can be incorporated into the weight-of-evidence against the suspect by defining the **match probability** for individual i, which is the conditional probability that i has trait $\mathcal{T}$ given that the suspect has this trait. Then it can be shown that

$$\mathbb{P}(G|E) = \frac{1}{1+\sum_{i=1}^{N} r_i},$$

which is a decreasing function of each of the $r_i's$, i.e., the presence of relatives in the pool of possible suspects decreases the weight of the DNA evidence against the suspect. On the other hand, if the suspect has no close relatives in the population, then $r_i = p$ for all i and this formula reduces to that shown in Example 1.2.

Typing errors: Let ϵ_1 denote the probability that the trait of an individual is wrongly recorded as $\mathcal{T}$ either through technical or clerical error, i.e., ϵ_1 is the probability of a false-positive. Then the probability that the suspect is guilty can be shown to be equal to

$$\mathbb{P}(G|E) = \frac{1}{1 + N(p + \epsilon_1)^2/p},$$

which is a decreasing function of ϵ_1. For example, if $p = 0.01$ and $\epsilon = 0.005$, then, $\mathbb{P}(G|E) = 1/3.25 \approx 0.31$. Here we have neglected the probability of a false-negative, but later we will show that this usually has a much smaller effect on $\mathbb{P}(G|E)$ whenever p is small.

Database searches: Now suppose that the suspect was identified after k other islanders were eliminated as suspects by being shown to not have the trait $\mathcal{T}$. This could happen, for example, if the traits of the suspect and these k other islanders had been recorded in a database that was searched in order to identify individuals with $\mathcal{T}$ Assuming that inclusion in the database is itself not culpatory, does the fact that the suspect was identified following a database search weaken or strengthen the evidence against that individual? Many scholars have argued that the evidence is weaker because the investigators set out to find a suspect with the trait. However, we will show that the contrary is true, i.e., the posterior probability is actually an increasing function of k:

$$\mathbb{P}(G|E) = \frac{1}{1 + (N - k)p}.$$

The reason for this result is that, in addition to the knowledge that the suspect has trait $\mathcal{T}$, we also know that k individuals living on the island do not have the trait and can eliminate these individuals as suspects. In particular, if $k = N$, then (ignoring typing error) all individuals except the suspect will have been eliminated from suspicion, which means that the suspect must be the guilty party.

Additional evidence: The previous calculations have assumed that in the absence of information on their $\mathcal{T}$-status, all individuals are equally likely to be guilty. In practice, there is usually additional evidence relevant to the question of guilt, e.g., criminal history, where the individuals live in relation to the crime scene, whether they have verifiable alibis, etc., which needs to be accounted for when calculating the probability that the suspect is guilty. This additional evidence can be summarized by a number w_i which denotes the weight of the non-$\mathcal{T}$ evidence against person i relative to its weight against the suspect. When $w_i > 1$, then i is more likely to be the culprit than the suspect given all of the information other than the fact that the culprit and the suspect both have trait $\mathcal{T}$. Then the probability that the suspect is guilty given all of the evidence is

$$\mathbb{P}(G|E) = \frac{1}{1 + p \times \sum_{i=1}^{N} w_i}.$$

If no additional evidence is available, then $w_i = 1$ for all $i = 1, \cdots, N$ and this formula reduces to that shown in Example 2.1.

Chapter 2

Probability Theory

2.1 Definitions and Examples

Our goal in this section is to develop a mathematical theory that can be used to analyze experiments with random outcomes. This structure will need to contain several components. We begin by defining the **sample space** to be the set of all possible outcomes. The sample space is often represented by the symbol Ω, although other symbols such as S may also be used. For example, if we roll a standard six-sided die, then the sample space could be taken to be $\Omega = \{1, 2, 3, 4, 5, 6\}$ since the die will land on a side displaying one of these numbers. However, we could also take the sample space to be the set of all natural numbers or even of all integers.

A subset E of Ω is said to be an **event** if the occurrence or non-occurrence of E can be deduced from our knowledge of the outcome of the experiment. Continuing with the previous example, if we are told which number is rolled, then every subset of Ω is an event since once we know which number is rolled, we will know whether that number is or is not an element of E. Alternatively, if we are only told whether the number rolled is even or odd, then the only subsets that will be events are the following: $\emptyset$, Ω, $\{1, 3, 5\}$, and $\{2, 4, 6\}$. In this case, $E = \{1\}$ is not an event because the evenness or oddness of the number rolled is not sufficient to tell us whether that number is equal to 1 or not. In this course, we will generally assume that all subsets of the sample space are events and we will use $\mathcal{F}$ to denote the collection of events.

Our third and final object is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ called the **probability distribution** on the sample space. Notice that $\mathbb{P}$ takes events (sets) as arguments and takes values between 0 and 1. If E is an event, then $\mathbb{P}(E)$ is a number between 0 and 1 that is called **the probability of the event** E. We can interpret the probability of an event in several ways. Perhaps the most familiar interpretation is that $\mathbb{P}(E)$ is the fraction of an infinite sequence of independent replicates of the experiment that result in an outcome that belongs to E, i.e., the fraction of trials in which the event E occurs. This is said to be the **frequentist interpretation of probability**. Alternatively, according to the **subjective interpretation of probability**, the probability of an event E is a measure of a person's belief in the likelihood that E will occur. This quantity can vary from person to person, but can be defined even when we cannot reasonably imagine repeating the experiment of interest.

Whichever interpretation we adopt, there are several properties that we will require of $\mathbb{P}$. First, because an experiment must result in some outcome that is contained in the sample space, we will insist that

$$\mathbb{P}(\emptyset) = 0 \quad \text{and} \quad \mathbb{P}(\Omega) = 1.$$

This, combined with the stipulation that $0 \leq \mathbb{P}(E) \leq 1$, is sometimes called the **first law of probability**. Next, let us say that two events E and F are **mutually exclusive** if E and F share no outcomes in common, i.e., $E \cap F = \emptyset$. Then the union $E \cup F$ is the event that either

E occurs or F occurs and the **second law of probability** asserts that

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F).$$

More generally, we say that the events $E_1, \cdots, E_n$ are disjoint if $E_i \cap E_j = \emptyset$ for any $1 \leq i \neq j \leq n$ and then we require that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i),$$

i.e., the probability of any finite union of disjoint events is equal to the sum of the probabilities of each of the events. This property is called **additivity** and also holds for countably infinite unions of mutually exclusive events.

Example 2.1. *We continue with the die example from above. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and let $\mathcal{F}$ be the collection of all subsets of Ω, i.e., $\mathcal{F}$ is the **power set** of the sample space. If the die is fair, then each outcome is equally likely and the first and second laws of probability imply that*

$$\mathbb{P}(\{i\}) = \frac{1}{6}, \quad 1 \leq i \leq 6,$$

since

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_{i=1}^{6} \{i\}\right) = \sum_{i=1}^{6} \mathbb{P}(\{i\}) = 6 \cdot \mathbb{P}\big(\{i\}\big)$$

for any $i \in \Omega$. Furthermore, if E is an event in Ω, then by the second law of probability, we have

$$\mathbb{P}(E) = \mathbb{P}\left(\bigcup_{i \in E} \{i\}\right) = |E| \times \frac{1}{6},$$

*where $|E|$ denotes the **cardinality** of E, i.e., the number of elements contained in E.*

Recall that the **complement** of a set $A \subset \Omega$ is the set of all points in Ω that are not in A:

$$A^c = \{x \in \Omega : x \notin A\}.$$

The next proposition outlines some simple identities and inequalities that are useful in many settings.

Proposition 2.1. *The following properties hold for any two events A, B in a probability space:*

1. *Complements: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*
2. *Subsets: if $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$;*
3. *Disjoint events: If A and B are mutually exclusive, then*

$$\mathbb{P}(A \cap B) = 0.$$

4. *Unions: For any two events A and B (not necessarily mutually exclusive), we have:*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Example 2.2. *Suppose that the frequency of HIV infection in a community is 0.05, that the frequency of tuberculosis infection is 0.1, and that the frequency of dual infection is 0.01. What proportion of individuals have neither HIV nor tuberculosis?*

$$\begin{aligned} \mathbb{P}(\textit{HIV or tuberculosis}) &= \mathbb{P}(\textit{HIV}) + \mathbb{P}(\textit{tuberculosis}) - \mathbb{P}(\textit{both}) = 0.05 + 0.1 - 0.01 = 0.14 \\ \mathbb{P}(\textit{neither}) &= 1 - \mathbb{P}(\textit{HIV or tuberculosis}) = 0.86 \end{aligned}$$

2.1.1 Independence

One of the most important concepts in probability theory is that of independence.

Definition 2.1. *Two events E and F contained in the same probability space are said to be* ***independent*** *if*

$$\mathbb{P}(E \cap F) = \mathbb{P}(E) \times \mathbb{P}(F).$$

The product rule that appears within the definition of independence is sometimes referred to as the **third law of probability.** In the next section we will see that two events are independent if the knowledge that one has occurred does not change the likelihood that the other has also occurred. However, we first consider some examples.

Example 2.3. *Suppose that a fair die is rolled twice and that the two rolls are independent. If the sample space is taken to be $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$, then the probability distribution can be defined by setting $\mathbb{P}((i,j)) = 1/36$, i.e., each one of the 36 outcomes in Ω is equally likely. To see that the numbers obtained on the first and second rolls are independent, observe that if E_i is the event that the first number rolled is i and F_j is the event that the second number rolled is j, then*

$$\mathbb{P}(E_i) = \mathbb{P}\left(\bigcup_{k=1}^{6}\{(i,k)\}\right) = \sum_{k=1}^{6}\mathbb{P}((i,k)) = \frac{1}{6}$$

and similarly $\mathbb{P}(F_j) = 1/6$. Since $E_i \cap F_j = \{(i,j)\}$, it follows that

$$\mathbb{P}\left(E_i \cap F_j\right) = \frac{1}{36} = \frac{1}{6}\cdot\frac{1}{6} = \mathbb{P}(E_i)\cdot\mathbb{P}(F_j),$$

which shows that E_i and F_j are independent.

Example 2.4. *Continue with the probability space introduced in the preceding example, but now let A be the event that the sum of the two numbers rolled is 7 and let B be the event that the second number rolled is 4. Then*

$$\mathbb{P}(A) = \mathbb{P}((1,6)) + \mathbb{P}((2,5)) + \cdots + \mathbb{P}((6,1)) = 6 \times \frac{1}{36} = \frac{1}{6}$$

and $\mathbb{P}(B) = 1/6$. Since $A \cap B = \{(3,4)\}$,

$$\mathbb{P}(A \cap B) = \frac{1}{36} = \frac{1}{6}\cdot\frac{1}{6} = \mathbb{P}(A)\cdot\mathbb{P}(B)$$

and it follows that A and B are independent.

Independence of collections containing more than two events is somewhat more complicated.

Definition 2.2. *The events $E_1, \cdots, E_n$ are said to be independent if the following identity holds for every set of indices $1 \leq i_1 < i_2 < \cdots < i_m \leq n$:*

$$\mathbb{P}\left(\bigcap_{j=1}^{m} E_{i_j}\right) = \prod_{j=1}^{m}\mathbb{P}\left(E_{i_j}\right).$$

In particular,

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \prod_{i=1}^{n}\mathbb{P}(E_i)$$

and

$$\mathbb{P}(E_i \cap E_j) = \mathbb{P}(E_i)\cdot\mathbb{P}(E_j)$$

for all $1 \leq i < j \leq n$.

Example 2.5. *Suppose that* $\boldsymbol{A/a}$, $\boldsymbol{B/b}$ *and* $\boldsymbol{C/c}$ *are alleles segregating at three unlinked loci. What is the probability that the child of two individuals that are heterozygous at all three loci is itself homozygous at at least one of these loci?*

The answer can be found most easily calculating the probability that the offspring is a triple heterozygote. Since the loci are unlinked, Mendel's law of independent assortment tells us that the offspring genotypes at the three loci are independent of one another, so that

$$\mathbb{P}(\textit{offspring is } \boldsymbol{AaBbCc}) = \mathbb{P}(\boldsymbol{Aa}) \times \mathbb{P}(\boldsymbol{Bb}) \times \mathbb{P}(\boldsymbol{Cc}).$$

To calculate the probability that the offspring is heterozygous at the first locus, recall that Mendel's law of segregation states that each parent contributes one of the two alleles and that each of the two alleles present in a parent is equally likely to be transmitted. Thus the offspring will have the $\boldsymbol{Aa}$ *genotype if it inherits* $\boldsymbol{A}$ *from its mother and* $\boldsymbol{a}$ *from its father or if it inherits* $\boldsymbol{a}$ *from its mother and* $\boldsymbol{A}$ *from its father. Assuming that the alleles transmitted by each parent are independent of one another, this implies that*

$$\mathbb{P}(\boldsymbol{Aa}) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}.$$

Since the other two loci behave similarly, we have $\mathbb{P}(\boldsymbol{Bb}) = \mathbb{P}(\boldsymbol{Cc}) = 1/2$ *and so* $\mathbb{P}(\boldsymbol{AaBbCc}) = 1/8$*. Finally, by using the first identity from Proposition (2.1), it follows that the probability that the offspring is homozygous at at least one locus is* $1 - 1/8 = 7/8$*.*

Example 2.6. *The recombination rate between two loci depends on several variables, including the physical distance between the two loci (if they occur on the same chromosome). One commonly used proxy for the recombination rate is the* ***recombination fraction****, which is defined as the probability that a randomly sampled gamete is recombinant at the two loci. For example, if the recombination fraction between two loci is* c*, then an individual that is heterozygous at both loci, say with parental genotypes* $\boldsymbol{AB/ab}$*, will produce gametes with both parental and recombinant genotypes with the following expected proportions:*

$$\frac{1}{2}(1-c)\,\boldsymbol{AB} + \frac{1}{2}(1-c)\,\boldsymbol{ab} + \frac{1}{2}c\,\boldsymbol{Ab} + \frac{1}{2}c\,\boldsymbol{aB}.$$

In particular, if the two loci reside on separate chromosomes, then according to Mendel's law of independent assortment, the expected frequencies of the four genotypes will be

$$\frac{1}{4}\,\boldsymbol{AB} + \frac{1}{4}\,\boldsymbol{ab} + \frac{1}{4}\,\boldsymbol{Ab} + \frac{1}{4}\,\boldsymbol{aB}$$

which shows that the recombination fraction between two unlinked loci is $c = 1/2$*. In contrast, if* $c = 0$*, then all of the gametes will have parental genotypes (no recombination occurs) and the two loci are said to be completely linked. This is thought to be the case for loci on the human mitochondrial genome and on the non-recombining portion of the Y chromosome. In general, recombination fractions always lie in the interval* $[0, 1/2]$*.*

2.2 Conditional Probability

Suppose that two fair dice are rolled and that the sum of the two rolls is even. What is the probability that we have rolled a 1 with the first die?

Without the benefit of any theory, we could address this question empirically in the following way. Suppose that we perform this experiment N times, with N large, and let (x_i, y_i) be the

outcome obtained on the i'th trial, where x_i is the number rolled with the first die and y_i is the number rolled with the second die. Then we can estimate the probability of interest (which we will call P) by dividing the number of trials for which $x_i = 1$ and $x_i + y_i$ is even by the total number of trials that have $x_i + y_i$ even:

$$\begin{aligned} P &\approx \frac{\#\{i : x_i = 1 \text{ and } x_i + y_i \text{ even }\}}{\#\{i : x_i + y_i \text{ even }\}} \\ &= \frac{\#\{i : x_i = 1 \text{ and } x_i + y_i \text{ even }\}/N}{\#\{i : x_i + y_i \text{ even }\}/N}, \end{aligned}$$

where we have divided both the numerator and denominator by N in the second line. Now, notice that for N large, we expect the following approximations to hold:

$$\begin{aligned} \mathbb{P}\{\text{the first roll equals 1 and the sum is even}\} &\approx \#\{i : x_i = 1 \text{ and } x_i + y_i \text{ even }\}/N \\ \mathbb{P}\{\text{the sum is even}\} &\approx \#\{i : x_i + y_i \text{ even }\}/N, \end{aligned}$$

with these becoming identities as N tends to infinity. This suggests that we can evaluate the probability P exactly using the following equation:

$$P = \frac{\mathbb{P}\{\text{the first roll equals 1 and the sum is even}\}}{\mathbb{P}\{\text{the sum is even}\}} = \frac{3/36}{18/36} = \frac{1}{6}.$$

Here P is said to be the conditional probability that the first roll is equal to 1 given that the sum of the two rolls is even. This identity motivates the following definition.

Definition 2.3. *Let E and F be events and assume that $\mathbb{P}(F) > 0$. Then the* ***conditional probability of E given F*** *is defined by the ratio*

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

F can be thought of as some additional piece of information that we have concerning the outcome of an experiment. The conditional probability $\mathbb{P}(E|F)$ then summarizes how this additional information affects our belief that the event E occurs. Notice that if F and E are independent and $\mathbb{P}(F) > 0$ and $\mathbb{P}(E) > 0$, then $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$ and so

$$\begin{aligned} \mathbb{P}(E|F) &= \frac{\mathbb{P}(E)\mathbb{P}(F)}{\mathbb{P}(F)} = \mathbb{P}(E) \\ \mathbb{P}(F|E) &= \frac{\mathbb{P}(E)\mathbb{P}(F)}{\mathbb{P}(E)} = \mathbb{P}(F). \end{aligned}$$

In other words, if E and F are independent, then the fact that E occurs will not alter the probability that F occurs, and *vice versa*.

As can be deduced from the definition, conditional probabilities are themselves probabilities and thus inherit all of the properties that probabilities possess. In particular, the results of Proposition 2.1 apply to conditional probabilities and we can even condition on multiple pieces of information by conditioning on each piece of information sequentially. For example, if E, F, and G are events and we let $Q(E) = \mathbb{P}(E|F)$ denote the conditional probability of E given F, then the conditional probability of E given F and G is equal to

$$\mathbb{P}(E|F, G) = Q(E|G),$$

i.e., we can first condition on F and then condition on G. Furthermore, we will get the same result if instead we first condition on G and then condition on F, i.e., the order in which we

condition on additional pieces of information does not affect the final probability once all of the information has been accounted for.

The next result is a simple, but very useful consequence of the definition of conditional probability.

Proposition 2.2. ***(Multiplication Rule)*** *Suppose that E and F are events and that $\mathbb{P}(F) > 0$. Then*

$$\mathbb{P}(E \cap F) = \mathbb{P}(E|F) \cdot \mathbb{P}(F). \tag{2.1}$$

Similarly, if $E_1, \cdots, E_k$ are events and $\mathbb{P}(E_1 \cap \cdots \cap E_{k-1}) > 0$, then recursive application of equation (3.1) shows that

$$\mathbb{P}(E_1 \cap \cdots \cap E_k) = \mathbb{P}(E_1)\,\mathbb{P}(E_2|E_1)\,\mathbb{P}(E_3|E_1 \cap E_2) \cdots \mathbb{P}(E_k|E_1 \cap \cdots \cap E_{k-1}).$$

2.2.1 The Law of Total Probability

Let E and F be two events and notice that we can write E as the following **disjoint union:**

$$E = (E \cap F) \cup (E \cap F^c).$$

Then, using the additivity of probabilities and the definition of conditional probabilities, we obtain:

$$\begin{aligned} \mathbb{P}(E) &= \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c) \\ &= \mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)\mathbb{P}(F^c) \\ &= \mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)(1 - \mathbb{P}(F)). \end{aligned} \tag{2.2}$$

This formula can sometimes be used to calculate the probability of a complicated event by exploiting the additional information provided by knowing whether or not F has also occurred. The hard part is knowing how to choose F so that the conditional probabilities $\mathbb{P}(E|F)$ and $\mathbb{P}(E|F^c)$ are both easy to calculate. In fact, (2.1) can be generalized in the following manner.

Proposition 2.3. ***(Law of Total Probability)*** *Suppose that $F_1, \cdots, F_n$ are mutually exclusive events such that $\mathbb{P}(F_i) > 0$ for each $i = 1, \cdots, n$ and $E \subset F_1 \cup \cdots \cup F_n$. Then*

$$\begin{aligned} \mathbb{P}(E) &= \sum_{i=1}^{n} \mathbb{P}(E \cap F_i) \\ &= \sum_{i=1}^{n} \mathbb{P}(E|F_i) \cdot \mathbb{P}(F_i). \end{aligned}$$

2.3 Bayes' Formula

Bayes' formula describes the relationship between the two conditional probabilities $\mathbb{P}(E|F)$ and $\mathbb{P}(F|E)$.

Theorem 2.1. ***(Bayes' Formula)*** *Suppose that E and F are events such that $\mathbb{P}(E) > 0$. Then*

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(E|F) \cdot \mathbb{P}(F)}{\mathbb{P}(E)}. \tag{2.3}$$

Similarly, if $F_1, \cdots, F_n$ are events such that $\mathbb{P}(F_i) > 0$ for $i = 1, \cdots, n$, then

$$\mathbb{P}(F_j|E) = \frac{\mathbb{P}(E \cap F_j)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|F_j)\,\mathbb{P}(F_j)}{\sum_{i=1}^{n} \mathbb{P}(E|F_i)\,\mathbb{P}(F_i)}. \tag{2.4}$$

Example 2.7. *Assume that eye color is determined by an individual's genotype at a single locus segregating the alleles B and b, and that bb homozygotes have blue eyes, while both BB homozygotes and Bb heterozygotes have brown eyes. Here, brown eye color is said to be* ***dominant*** *over blue eye color. (In reality, the genetics of human eye color is more complicated, with multiple loci affecting this trait.) Suppose that Paul and his parents have brown eyes, and that Paul's sister has blue eyes.*

a) What is the probability that Paul is a Bb heterozygote?
b) Suppose that Paul's partner has blue eyes. What is the probability that their first child will have blue eyes?
c) Suppose that their first child has brown eyes. What is the probability that their second child will also have brown eyes?

Solutions:

a) Since Paul's sister has blue eyes, we know that her genotype must be bb and consequently both parents (who are brown-eyed) must be Bb heterozygotes. However, Paul's genotype could be BB or Bb. Let E_1 and E_2 be the events that Paul has genotype BB or Bb, respectively, and let F be the event that Paul has brown eyes. Then Mendel's law of segregation implies that

$$\mathbb{P}(F) = \mathbb{P}(E_1) + \mathbb{P}(E_2) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4},$$

while Bayes' formula can be used to calculate that

$$\mathbb{P}(E_2|F) \;=\; \frac{\mathbb{P}(F|E_2)\mathbb{P}(E_2)}{\mathbb{P}(F)} \;=\; \frac{1 \cdot 1/2}{3/4} \;=\; 2/3.$$

b) If Paul's partner has blue eyes, then her genotype is bb. Now if Paul has genotype BB, then his children can only have genotype Bb and therefore have zero probability of having blue eyes. On the other hand, if Paul has genotype Bb, then his first child will have blue eyes with probability $1/2$ (i.e., with the probability that the child inherits a b gene from Paul). Thus, the total probability that the first child has blue eyes is:

$$1/3 \cdot 0 + 2/3 \cdot 1/2 = 1/3.$$

c) Knowing that their first child has brown eyes provides us with additional information about the genotypes of the parents. Let G be the event that the first child has brown eyes and let E_1 and E_2 be the events that Paul has genotype BB or Bb, respectively. From (a) we know that $\mathbb{P}(E_1) = 1/3$ and $\mathbb{P}(E_2) = 2/3$, while (b) tells us that $\mathbb{P}(G) = 1 - 1/3 = 2/3$. Consequently,

$$\begin{aligned} \mathbb{P}(E_1|G) &= \frac{\mathbb{P}(G|E_1)\mathbb{P}(E_1)}{\mathbb{P}(G)} = \frac{1 \cdot 1/3}{2/3} = 1/2, \\ \mathbb{P}(E_2|G) &= 1 - \mathbb{P}(E_1|G) = 1/2. \end{aligned}$$

Let H be the event that the second child has brown eyes and notice that

$$\begin{aligned} \mathbb{P}(H|E_1, G) &= 1 \\ \mathbb{P}(H|E_2, G) &= 1/2. \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{P}(H|G) &= \mathbb{P}(H \cap E_1|G) + \mathbb{P}(H \cap E_2|G) \\ &= \mathbb{P}(H|E_1, G)\mathbb{P}(E_1|G) + \mathbb{P}(H|E_2, G)\mathbb{P}(E_2|G) \\ &= 1 \cdot 1/2 + 1/2 \cdot 1/2 = 3/4. \end{aligned}$$

2.4 Random Variables and Distributions

Our next definition is motivated by the following scenario. Suppose that an experiment is performed and that the random outcome is modeled by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Although a large amount of information may be available concerning the outcome ω, often we are primarily interested in a small number of numerical quantities that are fully determined by that outcome. For example, if an STR profile is generated from a blood sample, then we will usually be most interested in knowing the allelic composition of the sample which can be summarized by a list of numbers (the copy numbers at each locus). These numbers are examples of random variables.

Definition 2.4. *Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. A real-valued **random variable** is a function $X : \Omega \to \mathbb{R}$ with the property that*

$$X^{-1}(a, b) \equiv \{\omega \in \Omega : a < X(\omega) < b\} \in \mathcal{F}$$

for every pair of real numbers $-\infty < a < b < \infty$. In other words, X is a random variable if we can determine when the condition $a < X < b$ holds using the information available to us from the outcome of the experiment.

Although a random variable is always defined on a probability space, often this dependence is left implicit and we use the notation X to represent both the variable and the particular value that the variable assumes in a given experiment. In this case, we need to be able to specify the **distribution** of the random variable, which can be done in several ways. One object that uniquely determines the distribution of any real-valued random variable is the cumulative distribution function, which is defined as follows.

Definition 2.5. *If X is a real-valued random variable, then the **cumulative distribution function** (c.d.f.) of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by the formula*

$$F_X(x) \equiv \mathbb{P}(X \leq x).$$

It follows from the properties of probability distributions that any cumulative distribution function is non-decreasing, is right-continuous (but not necessarily continuous), and has the following two limits:

$$0 = \lim_{x \to -\infty} F_X(x) < \lim_{x \to \infty} F_X(x) = 1.$$

Furthermore, two random variables X and Y have the same distribution if and only if they have the same cumulative distribution functions, i.e., if and only if $F_X(x) = F_Y(x)$ for all $x \in (-\infty, \infty)$. Notice, however, that two variables can have the same distribution without themselves being equal.

There are two classes of random variables that will be of particular interest to us: discrete random variables and continuous random variables.

2.4.1 Discrete Random Variables

Definition 2.6. *A real-valued random variable is said to be **discrete** if the variable can only take values in a finite or countably infinite set $E = \{x_1, x_2, \cdots\}$. In this case, the distribution of the variable is uniquely determined by its **probability mass function** $p_X : E \to [0, 1]$ defined by*

$$p_X(x) \equiv \mathbb{P}(X = x).$$

Indeed, once we know the probability mass function of a discrete variable X, we can calculate the probabilities of other events involving the variable using the formula

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x),$$

whenever $A \subset E$. We can also use the probability mass function to calculate the mean and the variance of a discrete random variable.

Definition 2.7. *Let X be a discrete random variable with probability mass function p_X. Then the **expected value** (also called the **mean**) of X is the quantity $\mathbb{E}[X]$ defined by*

$$\mathbb{E}[X] \equiv \sum_{x_i} p_X(x_i) \cdot x_i.$$

In other words, the expected value of a discrete random variable is just the weighted average of the values that the random variable can take where the weights are given by the probabilities of the individual values. An important property of expectations is that they are linear, i.e., if $X_1, \cdots, X_n$ are random variables and $c_1, \cdots, c_n$ are real numbers, then

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i],$$

and this identity holds even if the variables $X_1, \cdots, X_n$ are not independent.

Given a random variable X with values in a set E and a function $f : E \to \mathbb{R}$, the composition $Y = f(X)$ is itself a real-valued random variable. The next result can sometimes be used to calculate the expected values of such random variables without having to explicitly determine the distribution of Y.

Proposition 2.4. *If X is a discrete random variable with values in the set E and $f : E \to \mathbb{R}$ is a function defined on this set, then $Y = f(X)$ is a discrete random variable and*

$$\mathbb{E}[f(X)] = \sum_{x \in E} p_X(x) \cdot f(x).$$

For example, if X is a real-valued random variable, then the **variance** of X is the quantity

$$\begin{aligned} \mathrm{Var}(X) &\equiv \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] \\ &= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2, \end{aligned}$$

and Proposition 2.4 can be used to evaluate either expectation whenever X is discrete with probability mass function p_X. The variance of a random variable measures the extent to which the variable tends to deviate from its mean, i.e., if the variance is small, then typically the variable will take on values that are close to the mean.

There are several families of discrete distributions that are especially important in applications.

Definition 2.8. *A random variable X is said to be a **Bernoulli** random variable with parameter $p \in [0, 1]$ if X takes values in the set $\{0, 1\}$ with probability mass function*

$$\begin{aligned} p_X(0) &= 1 - p \\ p_X(1) &= p. \end{aligned}$$

Bernoulli random variables are often used to encode the results of an experiment that either results in a success (in which case we take $X = 1$) or in a failure (in which case $X = 0$). In fact, the Bernoulli distribution is a special case of the following more general family of distributions.

Definition 2.9. *A random variable X is said to have the **binomial distribution** with parameters $n \geq 1$ and $p \in [0, 1]$ if X takes values in the set $\{0, 1, \cdots, n\}$ with probability mass function*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, \cdots, n\}.$$

The binomial coefficient $\binom{n}{k}$ is the number of ways of choosing a subset of k elements from a set containing n elements and is equal to

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where the factorial function is defined by setting $n! = n(n-1)(n-2)\cdots 2 \cdot 1$ whenever $n \geq 1$ and $0! = 1$.

The binomial distribution typically arises in the following way. Suppose that a sequence of n independent but otherwise identical trials are performed and each trial results in a success with probability p and a failure otherwise. Then the total number of successes is a binomial random variable with parameters n and p. Notice that when $n = 1$, a binomial random variable is just a Bernoulli random variable. Using Proposition 2.4, it can be shown that the mean and the variance of a Bernoulli random variable with parameters n and p are

$$\begin{aligned} \mathbb{E}[X] &= np \\ \mathrm{Var}(X) &= np(1-p). \end{aligned}$$

Definition 2.10. *A random variable X is said to have the **Poisson distribution** with parameter $\lambda > 0$ if X takes values in the non-negative integers $\{0, 1, 2, \cdots\}$ with probability mass function*

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \geq 0.$$

The mean and the variance of a Poisson random variable with parameter λ are both equal to λ. The next theorem, which is sometimes called the **Law of Rare Events**, explains why the Poisson distribution is of interest.

Theorem 2.2. *For each $n \geq 1$, let $X_{n,i}, 1 \leq i \leq n$, be a collection of independent Bernoulli random variables and let $p_{n,i} = \mathbb{P}(X_{n,i} = 1)$. Suppose that the following two assumptions hold:*

$$\begin{aligned} \lim_{n\to\infty} \sum_{i=1}^{n} p_{n,i} &= \lambda < \infty \\ \lim_{n\to\infty} \max_{1\leq i\leq n} p_{n,i} &= 0. \end{aligned}$$

If $X_n = X_{n,1} + \cdots + X_{n,n}$, then for any integer $k \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

The Law of Rare Events has the following interpretation. Suppose that a large number of independent trials is performed and let $p_{n,i}$ be the probability of success on the i'th trial. Here we

will not require all trials to be equally likely to be successful, but we will assume that each trial is individually unlikely to result in a success (hence $\max_{1\leq i\leq n} p_{n,i}$ is small). If $X_n = X_{n,1}+\cdots+X_{n,n}$ is the total number of successes in all n trials, then the Law of Rare Events implies that X_n is approximately Poisson distributed with parameter $\lambda = p_{n,1}+\cdots+p_{n,n}$, which is just the expected number of successes. For this reason, the Poisson distribution is often used to model the number of events that occur when the likelihood of any one event is rare and different events occur independently of one another. For example, the number of sites that differ between two allelic variants at a given locus in the genome is usually approximately Poisson distributed, because different sites mutate (approximately) independently of one another and because the probability of mutation at any particular site is typically very small.

2.4.2 Continuous Random Variables

Definition 2.11. *A real-valued random variable X is said to be **continuous** if there is a function $p_X : \mathbb{R} \to [0,\infty]$, called the **probability density function** of X, such that*

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p_X(x)dx$$

for every pair of numbers $a < b$.

An important difference between discrete and continuous random variables is that if X is continuous, then $\mathbb{P}(X = x) = 0$ for every $x \in (-\infty,\infty)$. Indeed,

$$\mathbb{P}(X = x) = \int_x^x p_X(y)dy = 0,$$

for every x and so the probability mass function of a continuous random variable is equal to zero everywhere. It follows that this function provides us with no information concerning the distribution of a continuous variable, which is why we must work with the probability density function instead. Notice that the probability density $p_X(x)$ is not itself a probability, since $p_X(x)$ can be greater than 1, but we can always calculate the probability of any event involving a continuous random variable by integrating its density over an appropriate subset of the real line, i.e.,

$$\mathbb{P}(X \in A) = \int_A p_X(x)dx$$

whenever $A \subset (-\infty,\infty)$. In particular, every density function integrates to 1 since

$$1 = \mathbb{P}(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} p_X(x)dx.$$

All of the definitions and propositions concerning expected values of discrete random variables have counterparts for continuous random variables, but now the probability mass function is replaced by the probability density and the sum is replaced by an integral.

Definition 2.12. *Suppose that X is continuous with probability density function p_X. Then the expected value of X is given by the formula*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} p_X(x)\cdot x dx.$$

Similarly, Proposition 2.4 has the following counterpart.

Proposition 2.5. *Suppose that X is continuous with probability density function p_X and let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function. Then*

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} p_X(x) f(x) dx.$$

The three most important families of continuous distributions are described below.

Definition 2.13. *Let $-\infty < a < b < \infty$ be real numbers. A continuous random variable X is said to be **uniformly distributed** on the interval $[a, b]$ if the density function of X is*

$$p_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

*If $a = 0$ and $b = 1$, then X is said to be a **standard uniform random variable**.*

In other words, X is uniformly distributed on $[a, b]$ if X only takes values in this interval and if all values in this interval are equally likely. In this case, Proposition 2.5 can be used to show that the mean and the variance of X are equal to

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Definition 2.14. *A continuous random variable X is said to be **exponentially distributed** with parameter $\lambda > 0$ if X takes values in the non-negative real numbers with density function*

$$p_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

The exponential distribution is often used to model the waiting time between events that occur at a constant rate λ, e.g., the time between successive mutations at a particular site in the genome. Notice that as $\lambda > 0$ increases, the time until the next event tends to decrease and, indeed, both the mean and the variance of X are inversely related to λ:

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

Definition 2.15. *A continuous random variable X is said to be **normally distributed** with mean μ and variance $\sigma^2 > 0$ if the density of X is*

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

*If $\mu = 0$ and $\sigma^2 = 1$, then X is said to be a **standard normal random variable.***

The significance of the normal distribution is explained by the following result which is called the **central limit theorem**.

Theorem 2.3. *Suppose that $X_1, X_2, \cdots$ are independent, identically-distributed random variables with mean μ and finite variance $\sigma^2 < \infty$ and let $S_n = X_1 + \cdots + X_n$. Then, for every pair of real numbers $a < b$,*

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b \right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

In other words, when n is large, the distribution of an appropriately normalized sum of I.I.D. random variables is approximately normal.

For example, the fact that certain quantitative traits such as adult height or weight are approximately normally distributed within human populations (when stratified by gender) can be explained by the central limit theorem. Adult height is known to be highly heritable, with approximately 80% of the variance in this trait explained by heredity. Furthermore, it is known that there are many hundreds of polymorphic loci that affect an individual's stature, and that these act in an approximately additive fashion. That is, an individual's height can be written as a sum

$$H = \bar{H} + \sum_{l=1}^{L} (X_{l,m} + X_{l,f}) + \epsilon,$$

where $\bar{H}$ is the average adult height in the population of interest and $X_{l,f}$ and $X_{l,m}$ are the deviations from this mean attributable to the alleles that the individual inherits at the l'th locus from their mother and father, respectively. Finally, ϵ is the component of the individual's height that is due to random environmental factors such as childhood nutrition, disease, etc.

2.5 Random Vectors

In many cases, an experiment will give rise to several random variables $X_1, \cdots, X_n$ that describe different features of the outcome. For example, $X_1, \cdots, X_{13}$ could denote the number of distinct alleles detected at each of the thirteen loci in a standard CODIS profile. In such cases, we are interested not only in the distribution of each variable taken on its own, but also in the distribution of the vector of values that all n variables take in any particular run of the experiment. To this end, we can define a **random vector** $X = (X_1, \cdots, X_n)$ and we define the **joint distribution** of the variables to be the distribution of the random vector X. The joint distribution of the variables tells us how the values of any one subset of the variables depend on the values assumed by the remaining variables. Notice that if we know the joint distribution of a vector of random variables, then we can determine the distribution of any one of the variables taken on its own - this is called the **marginal distribution** of that variable - but the converse is not true. Even if we know the marginal distribution of each of the n variables, this is not sufficient information in general to reconstruct the joint distribution. One exception is when we know that the n variables are independent.

Definition 2.16. *The random variables $X_1, \cdots, X_n$ are said to be independent if for every collection of sets $E_1, \cdots, E_n \subset \mathbb{R}$ the following identity holds:*

$$\mathbb{P}(X_1 \in E_1, \cdots, X_n \in E_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in E_i).$$

Another case of special interest is when each of the variables $X_1, \cdots, X_n$ is discrete with values in then countable sets $E_1, \cdots, E_n$. Then the random vector $X = (X_1, \cdots, X_n)$ is discrete, with values in the product space

$$E = E_1 \times \cdots \times E_n = \{(x_1, \cdots, x_n) : x_i \in E_i\},$$

and the joint probability mass function of X is the function $p : E \to [0, 1]$ defined by

$$p(x_1, \cdots, x_n) = \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n).$$

Example 2.8. *If $X_1, \cdots, X_n$ are independent discrete random variables with marginal probability mass functions $p_1, \cdots, p_n$, respectively, then their joint probability mass is equal to the product of the marginal probability mass functions:*

$$p(x_1, \cdots, x_n) = \prod_{i=1}^{n} p_i(x_i).$$

The multinomial distribution is an example of a multivariate distribution for which the component variables are not independent.

Definition 2.17. *A random vector $(X_1, \cdots, X_m)$ is said to have the **multinomial distribution** with parameters $n \geq 1$ and $(p_1, \cdots, p_m)$ if $X_1, \cdots, X_m$ take values in the set $\{0, \cdots, n\}$ with joint probability mass function*

$$p(n_1, \cdots, n_m) = \binom{n}{n_1, \cdots, n_m} \prod_{i=1}^{m} p_i^{n_i}$$

provided that $n_1 + \cdots + n_m = n$. Here $p_1, \cdots, p_m$ are non-negative numbers subject to the condition $p_1 + \cdots + p_m = 1$ and

$$\binom{n}{n_1, \cdots, n_m} = \frac{n!}{n_1! \cdots n_m!}$$

is the number of ways of choosing m subsets of sizes $n_1, \cdots, n_m$ from a set of n objects.

The multinomial distribution arises in the following setting. Suppose that n independent experiments are conducted and that each experiment can result in any one of m possible outcomes. If the probability that an experiment results in the i'th outcome is p_i and X_i denotes the total number of experiments with this outcome, then the random vector $(X_1, \cdots, X_m)$ has the multinomial distribution with parameters n and $(p_1, \cdots, p_m)$.

Chapter 3

Topics in Population Genetics

3.1 Hardy-Weinberg Equilibrium

Suppose that k alleles $A_1, \cdots, A_n$ are segregating at an autosomal locus. Then there are $k(k+1)/2$ distinct diploid genotypes: k homozygous genotypes A_iA_i and $\binom{k}{2} = \frac{k(k-1)}{2}$ heterozygous genotypes A_iA_j. (Recall that the order in which we write the two alleles in a heterozygous genotype usually does not matter.) A basic question of interest in population genetics is how the frequencies of the diploid genotypes relate to the frequencies of the alleles. The answer, of course, depends on the biological details of the system, but a result known as **Hardy-Weinberg equilibrium** seems to hold, at least approximately, at many loci.

To derive the Hardy-Weinberg equilibrium, we will make the following assumptions:

- the population is infinite and has non-overlapping generations;
- the population is **panmictic** and **random mating**, so that mate choice is independent of an individual's genotype;
- the locus is **selectively neutral**, i.e., an individual's genotype does not affect the probability that they survive and reproduce or the number of offspring that they leave;
- segregation during meiosis adheres to Mendel's first law, i.e., there is no **meiotic drive**;
- there is no mutation at the locus.

Let p_{ij} denote the frequency of A_iA_j heterozygotes in the current generation and let

$$p_i = p_{ii} + \frac{1}{2}\sum_{j \neq i} p_{ij}$$

be the corresponding frequency of the allele A_i. Under the above assumptions, the frequency of the genotype A_iA_j in the next generation is equal to

$$p'_{ij} = \begin{cases} p_i^2 & \text{if } j = i \\ 2p_ip_j & \text{if } j \neq i. \end{cases} \tag{3.1}$$

Although this observation is the one that is most often associated with the phrase Hardy-Weinberg equilibrium, there is another equally important point, which is that the frequency of allele A_i following one round of random mating is

$$p'_i \;\; = \;\; p'_{ii} + \frac{1}{2}\sum_{j \neq i} p'_{ij}$$

$$
\begin{aligned}
&= p_i^2 + \frac{1}{2}\sum_{j\neq i} 2p_i p_j \\
&= p_i\left(\sum_{j=1}^{k} p_j\right) = p_i,
\end{aligned}
$$

where the final identity follows from the fact that the sum of the frequencies of all k alleles is equal to 1. In other words, under the above assumptions, the frequencies of the alleles are constant across generations and the frequencies of the diploid genotypes are constant after a single round of random mating, no matter what the initial frequencies are.

There are many factors that can cause genotype frequencies to deviate from the values expected under Hardy-Weinberg equilibrium (HWE) and arguably the most important of these (at least in the context of forensic genetics) is population structure. We will examine the consequences of non-random mating being in a later section, but we first consider two statistical procedures that can be used to test whether the genotype counts in a random sample of individuals drawn from a population are consistent with the hypothesis that that locus is at HWE. A statistical test is necessary because even if HWE holds exactly in the population, it probably won't hold in a random sample of individuals from the population.

3.1.1 Pearson's Goodness-of-Fit Test

Suppose that n individuals are sampled from a population and let n_{ij} be the number of individuals in the sample with genotype A_{ij}. Then we can estimate the genotype and allele frequencies as

$$
\begin{aligned}
\hat{p}_{ij} &= \frac{n_{ij}}{n} \\
\hat{p}_i &= \frac{1}{2n}\left(2n_{ii} + \sum_{j\neq i} n_{ij}\right),
\end{aligned}
$$

where the hat over the variable indicates that it is an estimate rather than the true value. Furthermore, if HWE holds in the population and we define

$$
\hat{E}_{ij} = \begin{cases} n\hat{p}_i^2 & \text{if } j = i \\ 2n\hat{p}_i\hat{p}_j & \text{if } j \neq i, \end{cases} \tag{3.2}
$$

then in a sufficiently large sample we would expect each of the differences

$$
n_{ii} - \hat{E}_{ij}
$$

between the observed and the expected numbers of A_{ij} individuals in the sample to be small. Here I have put hats over the numbers $\hat{E}_{ij}$ to indicate that these are only estimates of the expected counts since these depend on our estimates of the allele frequencies. To test whether these differences are sufficiently small, we calculate the following **goodness-of-fit** test statistic

$$
T = \sum_{i\leq j} \frac{\left(n_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}
$$

and compare the observed value of T with the distribution of values that would be expected under HWE. Since the exact distribution is complicated, it is customary to instead use a **large sample approximation** based on the central limit theorem. For this particular problem, it can be shown that in the limit of infinite sample size, the distribution of T is asymptotic to a χ^2-distribution with $k(k-1)/2$ degrees of freedom. A χ^2-distribution with d degrees of freedom

is defined to be the distribution of the sum of the squares of d independent standard normal random variables. For our problem, the number of degrees of freedom is determined by noting that there are $k(k+1)/2$ different genotypes and that we lose one degree of freedom since the sum of these frequencies is 1 and we lose an additional $k-1$ degrees of freedom by using the sample to estimate the allele frequencies, giving

$$d = \frac{k(k+1)}{2} - 1 - (k-1) = \frac{k(k-1)}{2}$$

as the total number of degrees of freedom for this test. To carry out this test in R, enter the following command

```
> 1 - pchisq(T, df = d)
```

where T is the value of the test statistic and d is the number of degrees of freedom. Since `pchisq(q, df)` gives the cumulative probability of values less than or equal to q, we need to subtract this probability from 1 to calculate the probability of obtaining a test statistic as large as or greater than the observed statistic.

A rule of thumb for the use of the χ^2-distribution is that most of the counts n_{ij} should be at least as large as 5 and none should be smaller than 2. When some of these counts are too small, the central limit theorem will not apply to the corresponding terms in T and then the χ^2-distribution may be too crude an approximation. In such cases, it is better to use the method described in the next section.

3.1.2 Fisher's Exact Test

If the number of alleles segregating at a locus is large (as will generally be the case with the STR loci used in forensics and relatedness testing), then it is likely that some of the possible diploid genotypes will be represented by very few individuals in the sample or even be missing altogether. When this is true, the large-sample asymptotics used to justify the use of the χ^2-distribution may not be sufficiently accurate and it is preferable to work with the exact sampling distribution of the test statistic. One way to do this is by using **Fisher's exact test**, which can be described explicitly for a locus segregating two alleles, say A and a.

Suppose that we sample n individuals and that the counts of the three diploid genotypes AA, Aa and aa in this sample are n_{AA}, n_{Aa} and n_{aa}, respectively. Then the number of A and a alleles present in the sample are given by

$$n_A = 2n_{AA} + n_{Aa} \quad \text{and} \quad n_a = 2n_{aa} + n_{Aa}. \tag{3.3}$$

Let the true (but unknown) allele frequencies be p_A and $p_a = 1 - p_A$ and assume that the genotype frequencies in the population are in HWE. To carry out Fisher's exact test, we begin by calculating the probability that a sample of n individuals contains the observed genotype counts conditional on it containing n_A A alleles and n_a a alleles:

$$\mathbb{P}\big(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a\big) = \frac{\mathbb{P}\big(n_{AA}, n_{Aa}, n_{aa}\big)}{\mathbb{P}\big(n_A, n_a\big)} \tag{3.4}$$

provided that the allele counts and the genotype counts are consistent, i.e., equation (3.3) holds. Since the number of A alleles in a sample containing n individuals from a population in which the frequency of A is p_A is binomially distributed with parameters $2n$ and p_A, it follows that the probability in the denominator is given by

$$\mathbb{P}\big(n_A, n_a\big) = \binom{2n}{n_A} p_A^{n_A} p_a^{n_a} = \frac{(2n)!}{n_A! n_a!} p_A^{n_A} p_a^{n_a}$$

whenever $2n = n_A + n_a$. Similarly, since the vector of genotype counts (n_{AA}, n_{Aa}, n_{aa}) is multinomially distributed with parameters n and (p_{AA}, p_{Aa}, p_{aa}), it follows that the probability in the numerator is

$$\begin{aligned}
\mathbb{P}\big(n_{AA}, n_{Aa}, n_{aa}\big) &= \binom{n}{n_{AA}, n_{Aa}, n_{aa}} p_{AA}^{n_{Aa}} p_{aa}^{naa} \\
&= \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} p_A^{2n_{AA}} (2p_A p_a)^{n_{Aa}} p_a^{2n_a} \\
&= \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} 2^{n_{Aa}} p_A^{n_A} p_a^{n_a},
\end{aligned}$$

where the identity of the first and second lines is a consequence of HWE and the identity of the second and third lines is a consequence of equation (3.3). Upon substituting these results into equation (3.4), the terms involving the allele frequencies cancel and we obtain the following probability

$$\mathbb{P}\big(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a\big) = \frac{2^{n_{Aa}} n! n_A! n_a!}{(2n)! n_{AA}! n_{Aa}! n_{aa}!}. \tag{3.5}$$

The p-value for Fisher's exact test is then calculated by summing the conditional probabilities of all possible outcomes that are either as or less probable than the observed outcome. Since the only outcomes that are relevant are those which satisfy $n_A = 2x_{AA} + x_{Aa}$ and $n_a = 2x_{aa} + x_{Aa}$, this computation can be done by exhaustive enumeration.

Fisher's exact test can also be extended to loci segregating more than two alleles. As in the biallelic case, we first calculate the probability of the genotype counts conditional on the observed numbers of alleles

$$\mathbb{P}\big(n_{A_i A_j}, 1 \le i \le j \le k | n_{A_i}, 1 \le i \le k\big)$$

and then sum the conditional probabilities of all possible outcomes that are at least as unlikely as the observed outcome. In practice, there may be too many possibilities to enumerate explicitly and Monte Carlo techniques are used instead which sample from the set of possible outcomes.

3.1.3 The Inbreeding Coefficient

An alternative approach is to generalize the model underlying Hardy-Weinberg equilibrium by allowing inbreeding. Suppose that the assumptions listed at the beginning of this section are modified so that with probability $1 - f$ mating is random and with probability f a gamete unites with another gamete carrying the same allele. f is called the **inbreeding coefficient** and is a measure of the tendency of individuals to mate with other individuals who are more closely related to themselves than to another random-chosen member of the population. If there are two alleles A and a segregating at the locus with frequencies p_A and p_a, then under inbreeding, the genotype frequencies after a single generation will be equal to

$$p_{AA} = (1 - f)p_A^2 + f p_A \tag{3.6}$$

$$p_{Aa} = 2(1 - f)p_A p_a \tag{3.7}$$

$$p_{aa} = (1 - f)p_a^2 + f p_A. \tag{3.8}$$

Notice that if $f = 0$, then there is no inbreeding and we are back to the conditions underlying HWE, while if $f = 1$, then all individuals are completely inbred and only homozygotes will be found in the population. Given a sample of n individuals in which the counts of the three genotypes are n_{AA}, n_{Aa}, and n_{aa}, the **likelihood function** $L(p_A; f)$ of the inbreeding coefficient f and the allele frequency p_A is defined to be equal to the probability of the data under those particular values of f and p_A. (Notice that p_A is determined by p_A since the two frequencies

sum to 1.) Assuming that we have sampled from an infinite population, this probability is given by the multinomial distribution with parameters n and (p_{AA}, p_{Aa}, p_{aa}) and so

$$\begin{aligned} L(p_A; f) &\equiv \binom{n}{n_{AA}, n_{Aa}, n_{aa}} p_{AA}^{n_{AA}} p_{Aa}^{n_{Aa}} p_{aa}^{n_{aa}} \\ &= \binom{n}{n_{AA}, n_{Aa}, n_{aa}} \left((1-f)p_A^2 + f p_A\right)^{n_{AA}} \left(2(1-f)p_A p_a\right)^{n_{Aa}} \left((1-f)p_a^2 + f p_A\right)^{n_{aa}}. \end{aligned}$$

If the allele frequency p_A were known exactly, the inbreeding coefficient could be estimated from the data by finding the value of f that maximizes the likelihood of the data. This estimate is called the **maximum likelihood estimate** (MLE) of f. Although the MLE can be found analytically by solving a cubic equation, in practice the solution is found numerically using Newton's method or a similar algorithm. For the more realistic scenario in which the allele frequencies are not known exactly, these can be jointly estimated along with the inbreeding coefficient by maximizing the likelihood function with respect to both unknown parameters. In this case, the estimates can only be found numerically.

3.1.4 The Wahlund Effect

The **Wahlund effect** refers to an apparent deficit of heterozygotes that is often observed in structured populations. We will illustrate this with the same biallelic model considered in the previous section. Suppose that a population is subdivided into d subpopulations of equal size and that the frequencies of the alleles A and a in the i'th subpopulation are p_i and $q_i = 1 - p_i$. If Hardy-Weinberg equilibrium holds in each subpopulation, then the frequencies of the genotypes AA, Aa and aa in the i'th population will be equal to p_i^2, $2p_iq_i$ and q_i^2, respectively. Furthermore, the population-wide frequencies of the alleles and genotypes can be calculated by averaging across subpopulations:

$$\bar{p} = \frac{1}{d}\sum_{i=1}^{d} p_i, \quad \bar{q} = 1 - \bar{p}$$

$$\bar{p}_{AA} = \frac{1}{d}\sum_{i=1}^{d} p_i^2, \quad \bar{p}_{Aa} = \frac{1}{d}\sum_{i=1}^{d} 2p_iq_i, \quad \bar{p}_{aa} = \frac{1}{d}\sum_{i=1}^{d} q_i^2.$$

However, it can be shown that unless the allele frequencies are the same in all of the subpopulations, HWE will not hold at the population level. For example, the difference between the observed and the expected frequencies of AA homozygotes in the population is

$$\begin{aligned} \bar{p}_{AA} - \bar{p}_A^2 &= \frac{1}{d}\sum_{i=1}^{d} p_i^2 - \left(\frac{1}{d}\sum_{i=1}^{d} p_i\right)^2 \\ &= \frac{1}{d}\sum_{i=1}^{d} (p_i - \bar{p})^2 \equiv V, \end{aligned}$$

where V is the variance of the frequency of allele A across the d subpopulations. In general, it can be shown that

$$\bar{p}_{AA} = \bar{p}_A^2 + V, \quad \bar{p}_{Aa} = 2\bar{p}\bar{q} - 2V, \quad \bar{p}_{aa} = \bar{p}_a^2 + V,$$

which demonstrates that population substructure leads to an apparent excess of homozygous genotypes and deficit of heterozygous genotypes. This is unsurprising since our assumption that HWE holds within each subpopulation is tantamount to assuming that individuals only mate with other members of their own subpopulation, which become inbred in comparison with the entire population. This is a potentially serious problem in forensic DNA analysis, since the

presence of population structure in human populations means that we cannot simply assume HWE when calculating match probabilities at diploid loci. Furthermore, although geographical separation is an important source of population structure (mating is more more likely to occur between individuals that live in the same region), structure can also arise from less evident factors such as ethnicity, religion, race, etc. Of course, the model described in this section is somewhat unrealistic in that it assumes that the subpopulations are completely reproductively isolated, but heterozygote deficiency will also occur when there is gene flow between the subpopulations.

3.2 Genetic Drift

Although one of the predictions of Hardy-Weinberg equilibrium is that allele frequencies are constant across generations, we would only expect this to hold approximately in a real population and then only over relatively short time scales. In fact, there are multiple processes that can cause allele frequencies to change, sometimes rapidly, including genetic drift, mutation, migration, and selection.

Recall that one of our assumptions in the section on HWE was that the population is infinite. This allowed us to equate sampling probabilities with frequencies and meant that we could predict the exact frequencies of the genotypes in each generation. However, because individual survival, reproduction and mating all depend to some extent on sequences of chance events, the frequency of an allele or genotype in a finite population will fluctuate between generations. For example, if the population contains only two haploid individuals with genotypes A and a and by chance the second individual dies before reproducing, then only A alleles will be present in the next generation and so the frequency of this allele will have increased from 0.5 to 1. This process of random fluctuations in allele frequencies is known as **genetic drift**.

Stochastic models have played an important role in studies of genetic drift and its evolutionary consequences and here we consider one of the best known of these models. The **Wright-Fisher model** is a discrete-time Markov chain that was introduced by Sewall Wright and R. A. Fisher in the 1930's. It makes the following assumptions:

- The population is of constant size N and has non-overlapping generations.
- Each individual alive in generation $t+1$ 'chooses' its parent uniformly at random and with replacement from among the N individuals alive in generation t.
- We consider a haploid locus segregating neutral alleles A and a.
- There is no mutation and each individual inherits their parent's genotype.

If p_t denotes the frequency of allele A in generation t, then the conditional distribution of Np_{t+1} given $p_t = p$ is binomial with parameters N and p, i.e., there are N individuals and the genotype of each one is determined by randomly choosing one of the N individuals alive in the previous generation, which will have genotype A with probability p and genotype a with probability $1-p$. It follows that the expected frequency of A in generation $t+1$ is the same as the frequency of A in the preceding generation,

$$\mathbb{E}[p_{t+1}|p_t = p] = p,$$

and so genetic drift leaves the expected values of the allele frequencies unchanged. On the other hand, p_{t+1} is a random quantity which need not be equal to p and the conditional variance of p_{t+1} given $p_t = p$ is

$$\text{Var}\,(p_{t+1}|p_t = p) = \frac{p(1-p)}{N}.$$

Notice that this variance is inversely proportional to the size of the population. This means that the fluctuations in allele frequencies are likely to be substantially larger in small populations than in large populations and so genetic drift will usually be more pronounced in smaller populations.

If new genotypes are not repeatedly introduced into the population by mutation or migration, then genetic drift will ultimately lead to the **fixation** of one of the alleles that was originally present. At this point, all genetic variation will have been lost from that population. The rate at which genetic variation is removed from the population by genetic drift can be quantified by calculating the expected 'heterozygosity' of the population and seeing how this changes from generation to generation. Let $H_t = 2p_t(1-p_t)$ be the heterozygosity in the t'th generation and observe that

$$\begin{aligned}\mathbb{E}[H_{t+1}|p_t = p] &= 2\mathbb{E}\Big[p_{t+1} - p_{t+1}^2 | p_t = p\Big] \\ &= 2p - 2p^2 - \frac{2p(1-p)}{N} \\ &= \left(1 - \frac{1}{N}\right) H_t.\end{aligned}$$

By repeatedly conditioning on allele frequencies in all of the preceding generations, it can be shown that

$$\mathbb{E}[H_{t+1}] = \left(1 - \frac{1}{N}\right)^{t+1} H_0.$$

In other words, over time, the expected heterozygosity of the population decreases geometrically by a factor of $(1 - 1/N)$. Furthermore, since this factor is itself an increasing function of N, we see that the expected heterozygosity will decline more rapidly in smaller populations than in larger populations.

3.2.1 Mutation-Drift Balance

We saw in the preceding section that genetic drift tends to reduce the amount of genetic variation present in a population. Despite this, most populations do harbor some variation which is maintained by the introduction of novel genotypes by mutation. In fact, over a period of many generations, these two opposing processes will tend to drive genetic variation to levels at which the rate at which variation is lost is balanced by the rate at which it is introduced by mutation. This state is known as **mutation-drift balance** and depends on both the population size and the mutation rate.

One quantity that can be used to quantify genetic variation in a population is the probability of **identity-by-descent**, which we denote F. Two chromosomes are said to be identical by descent (ibd) if they share the same genotype which they inherited (without mutation) from a common ancestor. (Two chromosomes can be identical in genotype without being ibd through convergent molecular evolution.) Suppose that the population size is $2N$ (N diploid individuals carry $2N$ chromosomes) and that the probability of mutation per generation is μ. If the population is at equilibrium, then the value of F in the present generation will be equal to the value in the previous generation. To calculate this value, observe that if two chromosomes are sampled at random and with replacement (to simplify the following argument), then they can be ibd either through sharing a common ancestor in the previous generation or by having distinct ancestors that are themselves ibd. Under the Wright-Fisher model, the former event occurs with probability $1/2N$, since each individual chooses its parent by random independent sampling. Alternatively, with probability $(1 - 1/2N)$, they do not share a common ancestor in the previous generation, in which case the two ancestors are ibd with probability F. Of course, in either case, the two individuals will only be ibd if they are not mutants, which is true with probability $(1-\mu)^2$ if we

assume that different lineages mutate independent. This leads to the following identity

$$F = \left\{ \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F \right\} (1-\mu)^2,$$

which can then be solved for F, giving

$$F = \frac{\frac{1}{2N}}{\frac{1}{2N} - 1 + \frac{1}{(1-\mu)^2}} \approx \frac{1}{1+4N\mu} \tag{3.9}$$

where the latter approximation is valid if $\mu \ll 1$ (as is usually the case). Equation (3.9) shows that the probability of identity-by-descent in a population will be small if the product $4N\mu$ is large, which in turn will be true whenever genetic drift is weak relative to mutation.

It is also possible to derive the stationary distribution of allele frequencies for some finite population models with mutation. If the Wright-Fisher model is modified so that allele A mutates to a with probability ν and a mutates to A with probability μ, then the stationary distribution of the frequency p of allele A can be approximated by the Beta distribution with density

$$\pi(p) = \frac{1}{\beta(4N\mu, 4N\nu)} p^{4N\mu - 1}(1-p)^{4N\nu - 1}.$$

The normalizing constant $\beta(a,b)$ in this expression is the Beta function defined by

$$\beta(a,b) \equiv \int_0^1 x^{a-1}(1-x)^{b-1} dx.$$

This density is bimodal when $4N\mu, 4N\nu \ll 1$, in which case mutation is much weaker than genetic drift and the frequency of A will usually be either very close to 1 or very close to 0, i.e., most of the time, one of the two alleles will be close to being fixed in the population. By exploiting certain properties of the Beta function, it can be shown that the average frequency of A and the average heterozygosity in a population at equilibrium are given by

$$\begin{aligned} \mathbb{E}[p] &= \frac{\mu}{\mu + \nu} \\ \mathbb{E}[2p(1-p)] &= \frac{2\mu\nu}{(\mu+\nu)(\mu+\nu+\frac{1}{4N})}. \end{aligned}$$

Whereas the mean frequency of A is independent of the population size, we see that the expected heterozygosity is an increasing function of N. This result is unsurprising since genetic drift is weaker in larger populations, which in turn means that larger population swill tend to retain more genetic variation (i.e., higher heterozygosity) than smaller populations.

3.2.2 Population Structure

Although stochastic models of structured populations are very complicated, a great deal has been learned about the effects of population structure on genetic variation and differentiation from the study of several highly simplified models. One of these is known as **Wright's island model**, which makes the following assumptions:

- The population is subdivided into a large number D of subpopulations.
- Each subpopulation is of constant size and contains N diploid individuals.
- Mating is random within subpopulations, but each offspring migrates to a new subpopulation with probability m.

- The newborn individual replaces one of the N individuals alive in the subpopulation where it settles.

Here m can be interpreted as a migration or dispersal rate: when $m = 0$, the subpopulations are completely isolated, while if $m = 1$, then the population is effectively panmictic.

Because each subpopulation is much smaller than the population as a whole, genetic drift has a much stronger impact on the allele frequencies within subpopulations than on the global allele frequencies. In particular, the subpopulation allele frequencies fluctuate more rapidly than do the global allele frequencies which therefore we treat as constant. (This phenomenon of having both rapidly and slowly varying quantities in the same model is known as separation-of-timescales.) Suppose that there are K alleles, $A_1, \cdots, A_K$, with global allele frequencies $p_1, \cdots, p_K$ that sum to 1. Then it can be shown that the equilibrium distribution of allele frequencies within a subpopulation can be approximated by the Dirichlet distribution with density

$$\pi(x_1, \cdots, x_K | p_1, \cdots, p_K) = \frac{\Gamma(\lambda)}{\prod_{k=1}^K \Gamma(\lambda p_k)} \prod_{k=1}^K x_k^{\lambda p_k - 1} \tag{3.10}$$

where

$$\begin{aligned} \Gamma(c) &= \int_0^\infty x^{c-1} e^{-x} dx \quad \text{(Gamma function)} \\ \lambda &= \frac{1}{\theta} - 1 = 4Nm. \end{aligned}$$

When $K = 2$, this distribution is just the Beta distribution introduced in the previous section, albeit with different parameters.

The coefficient θ is usually denoted F_{ST} in the population genetics literature and measures the degree of genetic differentiation between the subpopulations. When $\theta \approx 0$, the subpopulations are essentially undifferentiated and the local allele frequencies deviate from the global frequencies very little. In contrast, when $\theta \approx 1$, then the subpopulations are highly differentiated and typically only one of the K alleles will be present in each subpopulation. For the island model described above, the equilibrium value of θ is equal to

$$\theta = \frac{1}{1 + 4Nm},$$

which will be small whenever the product $4Nm$ is large. This is reminiscent of the result for the equilibrium probability of ibd in a model with genetic drift and mutation, except that in the present scenario it is migration rather than mutation that is responsible for introducing genetic variation into a subpopulation. Most estimates of θ for human populations lie between 0 and 0.05, indicating that these populations are only weakly genetically differentiated, either because they are of recent origin or because of ongoing gene flow.

3.2.3 Sampling Distributions

Suppose that n chromosomes are sampled at random from a population and typed at a diploid locus at which K distinct alleles $A_1, \cdots, A_K$ have been identified. If the frequencies of these alleles are known, say $p_1, \cdots, p_K$, then the the probability that our sample contains n_1 copies of A_1, n_2 copies of A_2, and so forth, is given by the multinomial distribution with parameters n and $(p_1, \cdots, p_K)$:

$$p(n_1, \cdots, n_K | p_1, \cdots, p_K) = \binom{n}{n_1, \cdots, n_K} \prod_{k=1}^K p_k^{n_k}. \tag{3.11}$$

This expression assumes that sampling is completely random with respect to any latent structure in the population, i.e., the probability that a particular group of chromosomes is sampled is independent of their membership in any genetically-differentiated subpopulations. In particular, if the sample is obtained by sampling individuals at random and typing both of the chromosomes that these individuals carry, then the multinomial distribution will only be appropriate when the population is at Hardy-Weinberg equilibrium. However, as we saw in Section 3.1.4, HWE usually will not hold in the population as a whole when there are subpopulations with different allele frequencies.

In some cases, this problem can be addressed with the help of the island model described in the previous section. Suppose that the population is subdivided into a large number of demes and that the unknown allele frequencies in each deme are distributed according to the Dirichlet distribution shown in Equation (3.10). Recall that (3.10) depends on the global allele frequencies, $p_1, \cdots, p_K$, which we assume are known. If these assumptions hold, then we can derive the sampling distribution of a group of chromosomes sampled from the same subpopulation by averaging the multinomial probability shown in (3.11) over the distribution of the unknown allele frequencies in that deme. To do so, we will need to integrate over the $(K-1)$-dimensional simplex $D_K = \{(x_1, \cdots, x_K) : x_1, \cdots, x_K \geq 0, x_1 + \cdots + x_K = 1\}$ and we obtain

$$\begin{aligned}
p_s(n_1, \cdots, n_K) &= \int_{D_K} p(n_1, \cdots, n_K | x_1, \cdots, x_K) \pi(x_1, \cdots, x_K | p_1, \cdots, p_K) dx_1 \cdots dx_K \\
&= \int_{D_K} \binom{n}{n_1, \cdots, n_K} \prod_{k=1}^{K} x_k^{n_k} \frac{\Gamma(\lambda)}{\prod_{i=1}^{K} \Gamma(\lambda p_i)} \prod_{k=1}^{K} x_k^{\lambda p_k - 1} dx_1 \cdots dx_K \\
&= \binom{n}{n_1, \cdots, n_K} \frac{\Gamma(\lambda)}{\prod_{k=1}^{K} \Gamma(\lambda p_k)} \int_{D_K} \prod_{k=1}^{K} x_k^{n_k + \lambda p_k - 1} dx_1 \cdots dx_K \\
&= \binom{n}{n_1, \cdots, n_K} \frac{\Gamma(\lambda)}{\prod_{k=1}^{K} \Gamma(\lambda p_k)} \frac{\prod_{k=1}^{K} \Gamma(\lambda p_k + n_k)}{\Gamma(\lambda + n)} \\
&= \binom{n}{n_1, \cdots, n_K} \frac{\Gamma(\lambda)}{\Gamma(\lambda + n)} \prod_{k=1}^{K} \frac{\Gamma(\lambda p_i + n_k)}{\Gamma(\lambda p_k)} \\
&= \binom{n}{n_1, \cdots, n_K} \left(\prod_{j=0}^{n-1} \frac{1}{\lambda + j} \right) \left(\prod_{k=1}^{K} \prod_{j=0}^{n_k - 1} (\lambda p_k + j) \right),
\end{aligned}$$

where we have used the following two identities to pass from the third to the fourth and from the fifth to the sixth lines in the above:

$$\begin{aligned}
&\int_{D_K} \prod_{k=1}^{K} x_k^{a_k - 1} dx_1 \cdots dx_K = \frac{\prod_{k=1}^{K} \Gamma(a_k)}{\Gamma(\sum_{k=1}^{K} a_k)} \\
&\Gamma(a + n) = (a + n - 1)\Gamma(a + n - 1) \quad \text{for any } a > 0.
\end{aligned}$$

The 's' subscript is included in p_s to remind us that this is the sampling distribution for a sample from a structured population. We can also express this distribution in terms of the kinship coefficient θ by substituting $\frac{1}{\theta} - 1$ for λ in the last line:

$$p_s(n_1, \cdots, n_K) = \binom{n}{n_1, \cdots, n_K} \frac{\prod_{k=1}^{K} \prod_{j=0}^{n_k - 1} (j\theta + (1-\theta)p_k)}{(1-\theta) \prod_{j=1}^{n-2} (1 + j\theta)}. \tag{3.12}$$

This is the probability of an unordered sample. In some instances, we will want to know the probability of an ordered sample in which A_1 appears n_1 times, etc., and this is obtained by

dividing (3.12) by the number of such ordered samples, which is given by the leading multinomial coefficient in that expression:

$$p_{s;o}(n_1, \cdots, n_K) = \frac{\prod_{k=1}^{K} \prod_{j=0}^{n_k-1} (j\theta + (1-\theta)p_k)}{(1-\theta)\prod_{j=1}^{n-2}(1+j\theta)}. \tag{3.13}$$

I have included an 'o' in the subscript of this probability to emphasize that it refers to an ordered sample. *Caveat:* Although Balding is careful to distinguish between ordered and unordered samples in his book, the two appear to be conflated in the discussion that appears on p. 39 of Fung & Hu (2008). In particular, equation (3.14) of the latter reference gives the probability of an ordered sample and is not equal to the line that immediately precedes it (which applies to an unordered sample).

Example 3.1. *If $n = n_i = 1$, then the only non-empty product in the numerator will be that corresponding to $k = i$ (empty products are set equal to 1 by convention) and we obtain*

$$p_s(A_i) = \frac{(1-\theta)p_i}{(1-\theta)} = p_i,$$

as expected since population structure does not change the sampling distribution of a single chromosome. If $n = 2$, then there are two possibilities. Suppose that both chromosomes carry A_i alleles, i.e., $n = n_i = 2$. Then

$$p_s(A_i, A_i) = \frac{(1-\theta)p_i \cdot (\theta + (1-\theta)p_i)}{(1-\theta)} = p_i^2 + \theta p_i(1-p_i).$$

Alternatively, if the sample contains one A_i allele and one A_j allele, where $i \neq j$, then

$$p_s(A_i, A_j) = 2\frac{(1-\theta)p_i \cdot (1-\theta)p_j}{(1-\theta)} = 2(1-\theta)p_i p_j,$$

where the factor of 2 accounts for the fact that the two alleles could be sampled in either order. Notice that when $\theta = 0$, these sampling probabilities are identical to those derived under HWE. However, when $\theta > 0$, there will be an excess of homozygous samples and a deficit of heterozygous samples, in accordance with the Wahlund effect.

Conditional Sampling Probabilities

One reason for writing down the ordered sample probability in (3.13) is so that we can use it to calculate the conditional distribution of alleles in future samples of chromosomes from a population that has already been sampled. For example, suppose that we previously sampled n chromosomes from a population and that n_1 of these were A_1 alleles, n_2 were A_2 alleles, etc. Then the conditional probability that the next chromosome sampled from the population will carry an A_i allele equal to

$$\frac{p_{s,o}(n_1, \cdots, n_i + 1, \cdots, n_K)}{p_{s,o}(n_1, \cdots, n_i, \cdots, n_K)} = \frac{n_i\theta + (1-\theta)p_i}{n\theta + (1-\theta)}. \tag{3.14}$$

Notice that this expression depends only on the number of A_i alleles in the original sample and on the size of that sample. For example, if we have previously sampled an A_1A_1 homozygote from a population, then the conditional distribution of the allele carried by a third chromosome that is sampled from the same population will be

$$\mathbb{P}(A_1|A_1A_1) \;=\; \frac{2\theta + (1-\theta)p_1}{2\theta + (1-\theta)} \;\geq\; p_1$$

$$\mathbb{P}(A_2|A_1A_1) \;=\; \frac{(1-\theta)p_2}{2\theta+(1-\theta)} \;\leq\; p_2,$$

and we will have strict inequalities on the right-hand side whenever $\theta > 0$. In other words, each time that we sample a given allele from a structured population, the chances of observing additional copies of that allele in future samples increase. The reason for this is that our initial sample is informative about the unknown frequencies of alleles in that particular subpopulation and our (posterior) estimates of these frequencies will either increase or decrease depending on whether the alleles are present in or absent from our original sample.

3.2.4 Linkage Equilibrium

A set of polymorphic loci is said to be in **linkage equilibrium** if the allelic states of the loci on a randomly sampled gamete are independent. For example, if there are n alleles $A_1, \cdots, A_n$ segregating at the first locus with frequencies $p_{A_1}, \cdots, p_{A_n}$ and m alleles $B_1, \cdots, B_m$ segregating at the second locus with frequencies $p_{B_1}, \cdots, q_{B_m}$, then the population will be at linkage equilibrium if the frequency of the two-locus gamete with alleles A_iB_j is

$$p_{A_iB_j} = p_{A_i}p_{B_j}$$

for each $1 \leq i \leq n$ and $1 \leq j \leq m$. If $n = m = 2$, then it suffices to check just one of these identities, say

$$p_{A_1B_1} = p_{A_1}p_{B_1},$$

as this will imply that the other three identities hold as well.

One prediction of Mendel's second law is that linkage equilibrium will hold at any set of neutral unlinked loci (e.g., loci located on different chromosomes) in an infinite randomly-mating population. In practice, these conditions never hold exactly and several factors, including physical linkage, genetic drift, population structure and selection, can create associations between the alleles found at different loci on the same gamete, a state known as linkage disequilibrium. For a two-locus, two-allele system, the strength of these associations can be quantified by a single statistic which is usually either the covariance between two of the allele frequencies

$$D = p_{A_1B_1} - p_{A_1}p_{B_1}$$

or the correlation coefficient

$$r = \frac{D}{\sqrt{p_{A_1}p_{A_2}p_{B_1}p_{B_2}}}.$$

Notice that $D = r = 0$ under linkage equilibrium and that positive values of D (or r) indicate that the alleles A_1 and B_1 occur together on the same gamete disproportionately often. Several methods, including Pearson's goodness-of-fit test and Fisher's exact test, have been developed that can be used to test for linkage equilibrium using samples of either haploid or (unphased) diploid genotypes. Although weak deviations can be found for some of the CODIS core STR loci, these are thought to mainly be due to population structure and are handled by using the sampling formula corrections described in the preceding section.

Chapter 4

DNA Profiling

4.1 Match Probabilities

The main statistical issue that must be addressed in genetic profiling is the following. Suppose that some DNA is recovered from a crime scene and is found to have the same genotype $\mathcal{G}$ as a sample of DNA collected from a known individual S. How strong is the evidence that S is also the source of the first sample? To address this question, we can set up a hypothesis test with the following hypotheses:

$$H_S : S \text{ is the source of the DNA in the first sample;}$$
$$H_I : \text{a different individual } I \text{ is the source of the first sample.}$$

Notice that the evidence E includes two pieces of information: that the source of the sample has genotype $\mathcal{G}$ and that individual S also has genotype $\mathcal{G}$. We can assess the relative support that the evidence provides for these two hypotheses by calculating the following likelihood ratio

$$L \equiv \frac{\mathbb{P}(E|H_S)}{\mathbb{P}(E|H_I)}.$$

In general, the larger the likelihood ratio L is, the stronger the evidence is in favor of H_S rather than H_I. If we assume that no errors have been made during the handling and genotyping of the data, then we can show that the reciprocal of the likelihood ratio is equal to the **match probability** of a random individual I to the profile $\mathcal{G}$, conditional on the observation that S also has this profile:

$$\begin{aligned} R_I \equiv L^{-1} &= \frac{\mathbb{P}(I \text{ and } S \text{ both have genotype } \mathcal{G})}{\mathbb{P}(S \text{ has genotype } \mathcal{G})} \\ &= \mathbb{P}\big(I \text{ has genotype } \mathcal{G} | S \text{ has genotype } \mathcal{G}\big). \end{aligned}$$

To make further progress, we need to be more specific about the genetic profile $\mathcal{G}$ and the population genetics of the relevant populations. Let us begin by considering a single diploid locus segregating alleles $A_1, \cdots, A_K$ at frequencies $p_1, \cdots, p_K$ and assume that S, I and the unknown source of the sample all come from the same population, which is at Hardy-Weinberg equilibrium. If S and I are unrelated, then the genotypes of S and I are independent and the match probability is just the unconditional probability that I has genotype $\mathcal{G} = A_i A_j$:

$$R_I = \begin{cases} p_i^2 & \text{if } j = i \\ 2p_i p_j & \text{if } j \neq i, \end{cases}$$

As expected, rarer genotypes $\mathcal{G}$ provide stronger evidence in favor of H_S.

Next, suppose that HWE is violated because of the presence of population structure. If we have no reason to believe that S and I belong to the same subpopulation, then we can continue to assume that the two are unrelated and therefore that their genotypes are independent. In the absence of specific information concerning the allele frequencies in the subpopulations, we can use the sampling distribution derived for the island model in Section 3.2.3 to calculate the match probability for I to genotype $\mathcal{G}$:

$$R_I = \begin{cases} (1-\theta)p_i^2 + \theta p_i & \text{if } j = i \\ 2(1-\theta)p_i p_j & \text{if } j \neq i, \end{cases}$$

where θ is the inbreeding coefficient. Since θ typically lies between 0 and 0.05 for large human populations, this correction will have only a very small impact on the likelihood ratio L when the sample genotype $\mathcal{G}$ is heterozygous. On the other hand, if the sample genotype is homozygous, say $\mathcal{G} = A_iA_i$, and the global frequency of the allele A_i is of the same order as θ, then inbreeding could reduce the likelihood ratio by a factor of two or more relative to the value calculated assuming HWE. In fact, since these expressions are based on a very crude model of human population structure and depend on a parameter θ that can only be estimated, the second National Research Council report (NRC II, 1996) recommended that match probabilities be calculated using the following formulas

$$R_I = \begin{cases} (1-\theta)p_i^2 + \theta p_i & \text{if } j = i \\ 2p_i p_j & \text{if } j \neq i. \end{cases}$$

These can be said to be conservative with respect to the hypothesis H_S in the sense that they probably underestimate the evidence in favor of H_S and thus would give a defendant the 'benefit of the doubt' in a criminal case.

If S and I belong to the same subpopulation, then the two will be more likely to share the same genotype than two individuals sampled at random from the population as a whole. In this case, the match probability can be calculated with the help of the conditional sampling distribution shown in Equation (3.14). First consider the case where $\mathcal{G} = A_iA_i$ is homozygous. Then the match probability is the same as the conditional probability that two additional A_i alleles are sampled given that two have already been sampled:

$$\begin{aligned} R_I &= \mathbb{P}(A_iA_i|A_iA_i) \\ &= \mathbb{P}(A_i|A_iA_i) \cdot \mathbb{P}(A_i|A_iA_iA_i) \\ &= \left(\frac{2\theta + (1-\theta)p_i}{1+\theta}\right)\left(\frac{3\theta + (1-\theta)p_i}{1+2\theta}\right) \\ &= \frac{(2\theta + (1-\theta)p_i)(3\theta + (1-\theta)p_i)}{(1+\theta)(1+2\theta)}. \end{aligned}$$

If instead the sample genotype $\mathcal{G} = A_iA_j$ is heterozygous, then we need to multiply the conditional sampling probability by a factor of two to account for the two different orders in which the two alleles of I could be sampled (this does not matter for S since the factor of 2 would appear in both the numerator and denominator of the match probability and thus will cancel). Repeated application of (3.14) here gives:

$$\begin{aligned} R_I &= 2 \cdot \mathbb{P}(A_iA_j|A_iA_j) \\ &= 2 \cdot \mathbb{P}(A_i|A_iA_j) \cdot \mathbb{P}(A_j|A_iA_iA_j) \\ &= 2 \cdot \left(\frac{\theta + (1-\theta)p_i}{1+\theta}\right)\left(\frac{\theta + (1-\theta)p_j}{1+2\theta}\right) \\ &= 2 \cdot \frac{(\theta + (1-\theta)p_i)(\theta + (1-\theta)p_j)}{(1+\theta)(1+2\theta)}. \end{aligned}$$

As a general rule, the match probabilities for homozygous genotypes are always increasing functions of θ, whereas the match probabilities for heterozygous genotypes can be either increasing or decreasing depending on the global allele frequencies. This is because inbreeding has two opposing effects on the conditional probability that a second individual is also a heterozygote. On the one hand, inbreeding reduces heterozygosity since individuals are disproportionately likely to inherit alleles that are identical-by-descent. On the other hand, if the global frequencies of A_i and A_j are small, then observing an individual with genotype A_iA_j in a subpopulation has the effect of shifting the conditional distribution of the local frequencies of the two alleles in that subpopulation to larger values, and this effect becomes stronger as θ increases. Since forensic markers are usually deliberately chosen so that the frequencies of the alleles are small (and therefore more informative about identity), the match probabilities for both kinds of genotypes are expected to be increasing functions of θ in practice.

4.1.1 Power of Discrimination

Although there are many polymorphic loci within the human genome, these are not all equally useful for the purposes of identification. One statistic that is sometimes used to evaluate the suitability of different loci for DNA profiling is the **power of discrimination** (PD), which is just the probability that two individuals, say X and Y, chosen at random have different genotypes at that locus. For example, the power of discrimination at a locus segregating K alleles $A_1, \cdots, A_K$ with frequencies $p_1, \cdots, p_K$ in a population that is panmictic and at Hardy-Weinberg equilibrium is equal to

$$\begin{aligned} PD &= 1 - \sum_{i \leq j} \mathbb{P}(X = A_iA_j, Y = A_iA_j) \\ &= 1 - \sum_{i \leq j} \mathbb{P}(X = A_iA_j)^2 \\ &= 1 - \left(\sum_{i=1}^{K} p_i^4 + 4 \sum_{i<j} p_i^2 p_j^2 \right). \end{aligned}$$

In general, the power of discrimination will be greatest at loci segregating many alleles at approximately equal frequencies.

4.1.2 Multiple Loci

Suppose that S is typed at several loci and found to have alleles $A_i^{(l)}A_j^{(l)}$ at the l'th locus, where $1 \leq l \leq L$. If the loci are in linkage equilibrium (cf. section 3.2.4), then the match probability of an individual I to the multilocus genotype $\mathcal{G}$ can be calculated using the **product rule**

$$R_I(\mathcal{G}) = \prod_{l=1}^{L} R_I^{(l)},$$

where $R_I^{(l)}$ is the single locus match probability at locus l given by one of the formulas derived above. This rule is expected to give a good approximation to the true match probability when the population is large and panmictic and when the different loci are unlinked, e.g., on different chromosomes, and evolve neutrally. Unfortunately, although three out of four of these conditions are more or less satisfied by the core STR loci used in forensic work, the fact that human populations are not panmictic is thought to be responsible for the modest levels of linkage disequilibrium (LD) that have been documented between some of these loci. This matters in

DNA profiling because one of the consequences of LD is to cause the match probabilities for common multi-locus genotypes to be underestimated when the product rule is applied to single locus match probabilities based on global allele frequencies. This can then lead to overestimates of the strength of the evidence indicating that a suspect S is the true source of a sample of DNA. Although one could correct for this bias by formulating a multi-locus extension of the island model, a simpler solution is to apply the product rule to one-locus match probabilities that have been calculated using overestimates of the inbreeding coefficient θ. For example, if the 'true' value of θ is estimated to be approximately 0.01, one takes $\theta = 0.02$ when computing $R_I^{(l)}$ and then multiplies these match probabilities together. Balding (2005) suggests that this approach probably overestimates the multi-locus match probabilities and therefore errs in favor of the suspect S.

When the product rule holds, the power of discrimination for a set of loci can be calculated using the formula

$$PD = 1 - \prod_{l=1}^{L}(1 - PD_l)$$

where PD_l is the probability of discrimination at the l'th locus in the set. For example, if $PD_l = 0.9$ for each $l = 1, \cdots, 13$ (a typical value for the core CODIS STR loci), then the power of discrimination for the complete 13-locus genotype is

$$PD = 1 - \left(1 - \frac{9}{10}\right)^{13} = 1 - \times 10^{-13}.$$

Of course, the actual match probabilities for specific genotypes can be substantially larger or smaller than the difference $1 - PD$, so the probability of discrimination only measures how well the set of markers performs on average.

4.2 Relatives

To calculate the match probability when I is a close relative of S, let Z be the number of alleles that are identical by descent (ibd) in I and S at the typed locus and define the **relatedness coefficients** $(k_0, 2k_1, k_2)$ as

$$\begin{aligned} k_0 &= \mathbb{P}(Z = 0) \\ 2k_1 &= \mathbb{P}(Z = 1) \\ k_2 &= \mathbb{P}(Z = 2). \end{aligned}$$

For example, in the absence of inbreeding, a parent and child will share exactly one allele that is ibd and so $k_0 = k_2 = 0$ and $2k_1 = 1$. Similarly, if I and S are full sibs and their parents are unrelated, then repeated application of Mendel's first law shows that $k_0 = k_2 = 1/4$ and $2k_1 = 1/2$. In general, the more closely related two individuals are, the larger will be the values of k_1 and k_2. In fact, we can combine the relatedness coefficients into a single number F, known as the **kinship coefficient**, that is a simple measure of the genetic relatedness of two individuals. F is defined to be the probability that two alleles, one sampled at random from each individual I and S, are identical by descent. Furthermore, it can be shown that F is the average of k_1 and k_2:

$$F = \frac{1}{2}(k_1 + k_2).$$

Caveat: Unfortunately, Balding (2005) and Fung & Hu (2008) use different conventions in their discussions of relatedness. My discussion follows section 3.6 of the latter reference because I will need it for material that we will cover later in the course, but I actually prefer Balding's notation

(section 6.2.4) which defines probabilities $\kappa_t = \mathbb{P}(Z = t)$ for $t = 0, 1, 2$. For comparison, notice that $\kappa_0 = k_0$, $\kappa_2 = k_2$ and $\kappa_1 = 2k_1$.

By conditioning on the number of alleles that are ibd between I and S and using the law of total probability, the match probability of I to $\mathcal{G}$ can be written as

$$\begin{aligned} R_I &= \sum_{t=0}^{2} \mathbb{P}\big(I = \mathcal{G} | S = \mathcal{G},\, Z = t\big) \cdot \mathbb{P}(Z = t) \\ &\equiv m_0 \cdot k_0 + m_1 \cdot (2k_1) + m_2 \cdot k_2, \end{aligned}$$

where m_t is the match probability for I to $\mathcal{G}$ when I and S share t alleles that are ibd at that locus. Notice that m_t depends on $\mathcal{G}$, whereas k_t does not. If $t = 2$, then I and S necessarily have the same genotype at that locus and so the conditional match probability is just $m_2 = 1$. If $t = 0$, then I and S share no alleles at the locus that are ibd, which means that the probability of a match for I to $\mathcal{G}$ is the same as that for an individual that is randomly chosen from the population (or subpopulation) and so m_0 can be calculated using the most appropriate formula from the preceding section. The case $t = 1$ is somewhat more complicated. If the genotype $\mathcal{G} = A_iA_i$ is homozygous, then I will necessarily have at least one A_i allele which is identical by descent to one of those carried by S and so m_1 is just the conditional probability that I has a second A_i allele conditional on S having genotype $\mathcal{G}$:

$$m_1 = \mathbb{P}(A_i | A_iA_i) = \frac{2\theta + (1-\theta)p_i}{1+\theta} \qquad \mathcal{G} = A_iA_i$$

If, instead, S has the heterozygous genotype $\mathcal{G} = A_iA_j$ with $j \neq i$, then the allele that is identical by descent in I and S is equally likely to be A_i or A_j. Taking these two possibilities into account leads to the following expression for m_1:

$$\begin{aligned} m_1 &= \frac{1}{2} \cdot \mathbb{P}(A_i | A_iA_j) + \frac{1}{2} \cdot \mathbb{P}(A_j | A_iA_j) \\ &= \frac{1}{2} \frac{(\theta + (1-\theta)p_i) + (\theta + (1-\theta)p_j)}{1+\theta} \\ &= \frac{\theta + (1-\theta)(p_i + p_j)/2}{1+\theta} \qquad \mathcal{G} = A_iA_j. \end{aligned}$$

Some explicit formulas for R_I are given in Table 3.15 of Fung & Hu (2008) for different relationships between I and S.

4.3 Additional Evidence

Suppose that there is additional evidence $\mathcal{E}$ (favorable or unfavorable) relevant to the proposition that an individual S is the source C of a sample of DNA and let w_i be the following likelihood ratio

$$w_i = \frac{\mathbb{P}(C = i | \mathcal{E})}{\mathbb{P}(C = S | \mathcal{E})},$$

where $C = i$ is the event that individual i is the source and $C = S$ is the event that S is the source. For instance, in a criminal investigation, additional evidence could come from witnesses, phone records, alibis, etc. To combine this evidence with the observation that S has the same genotype $\mathcal{G}$ as C, we use Bayes' formula

$$\begin{aligned} \mathbb{P}(C = S | \mathcal{G}, \mathcal{E}) &= \frac{\mathbb{P}(\mathcal{G} | C = S, \mathcal{E}) \mathbb{P}(C = S | \mathcal{E})}{\mathbb{P}(\mathcal{G} | \mathcal{E})} \\ &= \frac{\mathbb{P}(\mathcal{G} | C = S) \mathbb{P}(C = S | \mathcal{E})}{\mathbb{P}(\mathcal{G}, C = S | \mathcal{E}) + \sum_{i \neq S} \mathbb{P}(\mathcal{G}, C = i | \mathcal{E})} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathbb{P}(\mathcal{G}|C=S)\mathbb{P}(C=S|\mathcal{E})}{\mathbb{P}(\mathcal{G}|C=S,\mathcal{E})\mathbb{P}(C=S|\mathcal{E}) + \sum_{i\neq S}\mathbb{P}(\mathcal{G}|C=i,\mathcal{E})\mathbb{P}(C=i|\mathcal{E})} \\
&= \frac{\mathbb{P}(\mathcal{G}|C=S)\mathbb{P}(C=S|\mathcal{E})}{\mathbb{P}(\mathcal{G}|C=S)\mathbb{P}(C=S|\mathcal{E}) + \sum_{i\neq S}\mathbb{P}(\mathcal{G}|C=i)\mathbb{P}(C=i|\mathcal{E})} \\
&= \left(1 + \sum_{i\neq S}\frac{\mathbb{P}(C=i|\mathcal{E})}{\mathbb{P}(C=S|\mathcal{E})}\frac{\mathbb{P}(\mathcal{G}|C=i)}{\mathbb{P}(\mathcal{G}|C=S)}\right)^{-1} \\
&= \frac{1}{1+\sum_{i\neq S} w_i R_i},
\end{aligned}
$$

where the summation is taken over all individuals that could be the source of the DNA and

$$R_i = \frac{\mathbb{P}(\mathcal{G}|C=i)}{\mathbb{P}(\mathcal{G}|C=S)} = L^{-1}$$

is the match probability introduced in section 4.1. This is equation (3.3) in Balding (2005), who calls this the weight-of-evidence formula. If no additional evidence is available, then $w_i = 1$ for every $i \neq S$ since every individual is equally likely to be the source of the DNA and we deduce that

$$\mathbb{P}(C=S|\mathcal{G}) = \frac{1}{1+\sum_{i\neq S} R_i}.$$

Chapter 5

Parentage Testing

5.1 Paternity Testing

In addition to its use in identification, DNA profiling can also be used to determine when two individuals are close biological relatives. Perhaps the most common application of this sort is to paternity testing, where the aim is to establish whether one individual is the father of another. In the **standard trio problem**, the genotypes of the mother M, the child C, and the alleged father AF are known and the problem is to decide between the following two hypotheses

$$\begin{aligned} &H_p : AF \text{ is the father of } C; \\ &H_d : \text{the father is a random member of population } \mathcal{P}, \end{aligned}$$

where $\mathcal{P}$ is the population containing AF and all other males that could have fathered C. The evidence in this problem consists of the three genotypes M, C and AF, and the relative support that this evidence gives to H_p compared with H_d is usually expressed in terms of the **paternity index**, which is just the likelihood ratio

$$\begin{aligned} PI &\equiv \frac{\mathbb{P}(C, M, AF|H_p)}{\mathbb{P}(C, M, AF|H_d)} \\ &= \frac{\mathbb{P}(C|M, AF, H_p)}{\mathbb{P}(C|M, AF, H_d)} \cdot \frac{\mathbb{P}(M, AF|H_p)}{\mathbb{P}(M, AF|H_d)}. \end{aligned}$$

The product in the second line can be simplified if we assume that mating is random with respect to an individual's genotype at the markers used in the test. In this case, the genotypes of M and AF are independent of the two hypotheses H_p and H_d and so

$$\mathbb{P}(M, AF|H_p) = \mathbb{P}(M, AF) = \mathbb{P}(M, AF|H_d).$$

Likewise, because H_d posits that the true father I is a random member of $\mathcal{P}$, it follows that the genotypes of I and AF are conditionally independent given the genetic composition of $\mathcal{P}$ and so

$$\mathbb{P}(C|M, AF, H_d) = \mathbb{P}(C|M, H_d).$$

These identities show that the paternity index for the standard trio problem is equal to

$$PI = \frac{\mathbb{P}(C|M, AF, H_p)}{\mathbb{P}(C|M, H_d)}. \tag{5.1}$$

The probabilities in the numerator and denominator of the paternity index can be evaluated using Mendel's law of random segregation. Here we will assume that the population $\mathcal{P}$ is at Hardy-Weinberg equilibrium, although the sampling formulas derived using the island model in

section 3.2.3 could also be used if population structure is believed to be important. By way of illustration, suppose that a single locus is typed and that all three individuals M, C, and AF have genotype A_iA_i, where the frequency of A_i in $\mathcal{P}$ is p_i. Then

$$\mathbb{P}(C|M, AF, H_p) = 1,$$

since any offspring of M and AF must have genotype A_iA_i at this locus if mutation is neglected. On the other hand,

$$\mathbb{P}(C|M, H_d) = p_i,$$

since under the hypothesis of random mating, the allele transmitted to C by I is just a random draw from the population, which will be A_i with probability p_i. Thus the paternity index in this case is equal to $PI = 1/p_i$, which will be large when the frequency of A_i is small. For a slightly more complicated scenario, suppose that the three genotypes are $C = A_iA_j$, $M = A_iA_k$, and $AF = A_jA_l$, where $i \neq j \neq k \neq l$, i.e., C shares one allele in common with the mother and a different allele in common with the alleged father. In this case, Mendel's first law implies that

$$\mathbb{P}(C|M, AF, H_p) = \frac{1}{4},$$

while

$$\mathbb{P}(C|M, H_d) = \frac{1}{2}p_j,$$

and so the paternity index is equal to $1/(2p_j)$. Formulas for the other possible combinations of single locus genotypes are given in Table 4.1 of Fung & Hu (2008). If the individuals are genotyped at multiple loci that are at linkage equilibrium within $\mathcal{P}$, then the product rule can be used to express the multi-locus paternity index PI_{ml} in terms of the single locus indices PI_l:

$$PI_{ml} = \prod_{l=1}^{L} PI_l.$$

5.1.1 The true and alleged fathers are related

The likelihood ratio can also be used in paternity testing when it is suggested that the true father is a close relative of the alleged father. In this case the competing hypotheses are

$$\begin{aligned} H_p &: AF \text{ is the father of } C; \\ H_d^0 &: \text{the father is a relative of } AF, \end{aligned}$$

where H_d^0 specifies the relationship between AF and the true father, and the likelihood ratio is equal to

$$\begin{aligned} LR^0 &\equiv \frac{\mathbb{P}(C, M, AF|H_p)}{\mathbb{P}(C, M, AF|H_d^0)} \\ &= \frac{\mathbb{P}(C|M, AF, H_p)}{\mathbb{P}(C|M, AF, H_d^0)} \end{aligned} \tag{5.2}$$

under the assumption of random mating. Notice that the only difference between equations (5.1) and (5.2) is that the probability in the denominator of the likelihood ratio depends on the genotype of AF when the true father and the alleged father are close relatives. This is because the genotype of the true father is unknown under H_d^0 (we only have genetic profiles of M, C and AF) but is disproportionately likely to share alleles in common with AF when the two are closely related.

The denominator of (5.2) can be evaluated with the help of the law of total probability by conditioning on the number of alleles Z that are identical by descent between the alleged father and the child. To this end, let $k_0, 2k_1, k_2$ be the relatedness coefficients (cf. section 4.2) of AF and C under the hypothesis H_d^0. For example, if it is suggested that the true father and AF are full brothers, then AF is an uncle to C and so $k_0 = 1/2$, $2k_1 = 1/2$ and $k_2 = 0$ under the assumption that AF and M are unrelated. We will make this assumption throughout this section and so we will have $k_2 = 0$ in general, which in turn implies that $k_0 = 1 - 2k_1$ since $k_0 + 2k_1 + k_2 = 1$. We first observe that if AF is the true father of C, then $Z = 1$ necessarily and so

$$\mathbb{P}(C|M, AF, H_p) = \mathbb{P}(C|M, AF, Z = 1).$$

Next, by the law of total probability, we have

$$\begin{aligned}\mathbb{P}(C|M, AF, H_d^0) &= \mathbb{P}(C|M, AF, Z = 0, H_d^0) \cdot \mathbb{P}(Z = 0|H_d^0) + \\ &\quad \mathbb{P}(C|M, AF, Z = 1, H_d^0) \cdot \mathbb{P}(Z = 1|H_d^0) \\ &= \mathbb{P}(C|M, AF, Z = 0, H_d^0) \cdot (1 - 2k_1) + \\ &\quad \mathbb{P}(C|M, AF, Z = 1, H_d^0) \cdot (2k_1).\end{aligned}$$

If H_d^0 is true and $Z = 0$, then the allele transmitted by the true father is independent of the genotype of AF and so

$$\mathbb{P}(C|M, AF, Z = 0, H_d^0) = \mathbb{P}(C|M, H_d^0) = \mathbb{P}(C|M, H_d),$$

where the second identity follows from the fact that H_d^0 and H_d are equivalent hypotheses if the genotype of the alleged father AF is unknown. (Here we are only considering genetic evidence; if non-genetic evidence is available, i.e., concerning relationships between M and the relatives of AF, then these hypotheses may not be equivalent even when the genotype of AF is unknown.) Furthermore, if it is known that $Z = 1$, then the genotype of the child is conditionally independent of the genotype of the true father and so

$$\mathbb{P}(C|M, AF, Z = 1, H_d^0) = \mathbb{P}(C|M, AF, Z = 1) = \mathbb{P}(C|M, AF, H_p).$$

Substituting these identities into (5.2) and recalling the definition of PI in (5.1) shows that

$$\begin{aligned}LR^0 &= \frac{\mathbb{P}(C|M, AF, H_p)}{\mathbb{P}(C|M, H_d) \cdot (1 - 2k_1) + \mathbb{P}(C|M, AF, H_p) \cdot 2k_1} \\ &= \frac{1}{2k_1 + (1 - 2k_1)(PI)^{-1}}.\end{aligned}$$

It is also possible to express the likelihood ratio in terms of the kinship coefficient F between the alleged and true father implied by H_d^0. Recall that F is defined as the probability that two alleles, one sampled at random from each individual, are identical by descent. Since each of the two alleles of the true father are equally likely to be transmitted to the child, it follows that the probability $2k_1$ that AF and C share one allele that is ibd is equal to $2F$ or, equivalently, $k_1 = F$. Thus, the likelihood ratio can also be written as

$$LR^0 = \frac{1}{2F + (1 - 2F)(PI)^{-1}}. \tag{5.3}$$

Since this is a weighted harmonic mean of 1 and PI, we see that LR^0 will be closer to 1 than PI except when $LR^0 = PI = 1$. For example, if M, C and AF are all A_iA_i homozygotes, then $PI = 1/p_i$ and so

$$LR^0 = \frac{1}{2F + (1 - 2F)p_i}.$$

Similar expressions that cover other combinations of single-locus genotypes can be found in Table 4.1 of Fung & Hu (2008).

5.1.2 Mother unavailable

If the mother is unavailable for genotyping, then the evidence consists of the genotypes of the alleged father and the child. The likelihood ratio can still be used to decide between the following two hypotheses

$$H_p : AF \text{ is the father of } C;$$
$$H_d : \text{the father is a random member of population } \mathcal{P},$$

but now the paternity index is given by the formula

$$\begin{aligned} PI &= \frac{\mathbb{P}(C, AF|H_p)}{\mathbb{P}(C, AF|H_d)} \\ &= \frac{\mathbb{P}(C|AF, H_p)}{\mathbb{P}(C|AF, H_d)} \cdot \frac{\mathbb{P}(AF|H_p)}{\mathbb{P}(AF|H_d)} \\ &= \frac{\mathbb{P}(C|AF, H_p)}{\mathbb{P}(C|H_d)} \end{aligned}$$

assuming that mating is random and that I and AF are unrelated. In the following examples, we will assume that the population $\mathcal{P}$ is at Hardy-Weinberg equilibrium; when population structure is considered to be important, then the conditional sampling distribution for the island model can be used with a suitable choice of the inbreeding coefficient θ. If $C = AF = A_iA_i$, then the paternity index is equal to

$$PI = \frac{1 \cdot p_i}{p_i^2} = \frac{1}{p_i},$$

which agrees with the index that we calculated for the standard trio problem when all three individuals are A_iA_i homozygotes. On the other hand, if $C = A_iA_j$ and $AF = A_jA_l$, where $j \neq k \neq l$ then

$$PI = \frac{\frac{1}{2} \cdot p_i}{2p_ip_j} = \frac{1}{4p_j},$$

which is half the value of the paternity index when the mother is known to have genotype A_iA_k.

5.1.3 Alleged father unavailable

If the alleged father is unavailable for genotyping, paternity testing can still be conducted using the genotype of a close relative R of AF. In this case the competing hypotheses are

$$H_r : \text{the father is a close relative of } R;$$
$$H_d : \text{the father is a random member of population } \mathcal{P},$$

where H_r specifies the relationship between R and AF, and the likelihood ratio is sometimes called the **avuncular index**:

$$AI = \frac{\mathbb{P}(C, M, R|H_r)}{\mathbb{P}(C, M, R|H_d)} = \frac{\mathbb{P}(C|M, R, H_r)}{\mathbb{P}(C|M, H_d)}.$$

As in section 5.1.1, it suffices to specify the kinship coefficient F between R and AF. Although this likelihood ratio can be evaluated directly, it can also be expressed as the ratio of the paternity index for the standard trio problem and the likelihood ratio LR^0 for the alternative hypothesis that the true father is a close relative of AF with kinship coefficient F. This result depends on the following identity

$$\mathbb{P}(C|M, R, H_r) = \mathbb{P}(C|M, AF, H_d^0)$$

which holds because H_d^0 is equivalent to H_r when the roles (and genotypes) of R and AF are interchanged. Recalling equation (5.3), this implies that

$$\begin{aligned} AI &= \frac{\mathbb{P}(C|M, AF, H_d^0)}{\mathbb{P}(C|M, H_d)} \\ &= \frac{\mathbb{P}(C|M, AF, H_d^0)}{\mathbb{P}(C|M, AF, H_p)} \cdot \frac{\mathbb{P}(C|M, AF, H_p)}{\mathbb{P}(C|M, H_d)} \\ &= \frac{1}{LR^0} \cdot PI \\ &= (1 - 2F) + 2F \cdot PI, \end{aligned} \tag{5.4}$$

and so the avuncular index is a weighted arithmetic mean of 1 and PI. Notice that the effect of using a relatives' genotype to test for the paternity of AF is to shrink the likelihood ratio towards 1 by a factor of $1 - 2F$, which makes this test less powerful than the paternity test for the standard trio problem. A similar argument shows that equation (5.4) also holds when the maternal genotype is unavailable if the paternity index is calculated using the method described in section 5.1.2.

5.1.4 Determination of both parents

In certain cases, it may be necessary to determine whether an individual C is the child of a pair of individuals AM and AF. The choice of hypotheses to be tested depends on the details of the case, but the most common propositions are:

H_p : the alleged parents are the true parents of C;
H_d : the true parents are a random unrelated couple in $\mathcal{P}$.

The likelihood ratio can be evaluated as follows

$$\begin{aligned} LR &= \frac{\mathbb{P}(C, AM, AF|H_p)}{\mathbb{P}(C, AM, AF|H_d)} \\ &= \frac{\mathbb{P}(C|AM, AF, H_p)}{\mathbb{P}(C|AM, AF, H_d)} \cdot \frac{\mathbb{P}(AM, AF|H_p)}{\mathbb{P}(AM, AF|H_d)} \\ &= \frac{\mathbb{P}(C|AM, AF, H_p)}{\mathbb{P}(C|AM, AF, H_d)} \\ &= \frac{\mathbb{P}(C|AM, AF, H_p)}{\mathbb{P}(C)}. \end{aligned}$$

For example, if we assume that the population is at HWE, then if $C = A_iA_j$, $AM = A_iA_k$ and $AF = A_jA_l$, where $i \neq j \neq k$, then the likelihood ratio is

$$LR = \frac{\frac{1}{2} \cdot \frac{1}{2}}{2p_ip_j} = \frac{1}{8p_ip_j}.$$

As with the paternity index, multi-locus likelihood ratios are often calculated using the product rule, which assumes that the population is in linkage equilibrium.

5.2 Exclusion Probabilities

In addition to the paternity index, some laboratories report the probability of excluding a random man from paternity given the genotypes of the mother M and child C. If this probability is very

close to 1 and if the alleged father cannot be excluded on the basis of their genotype AF, then this provides support for the proposition that the alleged father is the true father. We first consider the exclusion probability based on a single autosomal locus segregating alleles $A_1, \cdots, A_k$, and assume that the population is at Hardy-Weinberg equilibrium. There are five separate cases, depending on the genotypes of the mother and child. If $C = A_iA_i$ and $M = A_iA_i$ or $M = A_iA_j$, then the child must have inherited an A_i allele from the father and so any man that lacks an A_i allele at this locus will be excluded as the father. It follows that the exclusion probability for a male sampled at random from this population is

$$EP = (1 - p_i)^2.$$

Similarly, if $C = A_iA_j$ and $M = A_iA_i$ or $M = A_iA_k$, then the child must have inherited an A_j allele from its father and so a random man will be excluded with probability

$$EP = (1 - p_j)^2.$$

Finally, if $C = M = A_iA_j$ with $j \neq i$, then the child must have inherited either an A_i or an A_j allele from its father and so the exclusion probability for a random many is no

$$EP = (1 - p_i - p_j)^2.$$

As expected, the rarer the alleles are that comprise the child's genotype, the greater will be the exclusion probability.

If multiple unlinked loci are typed (as will usually be the case) and linkage equilibrium approximately holds between these loci, then the overall probability of exclusion based on the multi-locus genotype is given by the following expression

$$EP_{ml} = 1 - \prod_{l=1}^{L}(1 - EP_l).$$

For example, if 10 unlinked loci are typed and the exclusion probability for each locus is $EP_l \approx 0.5$ say, then the overall exclusion probability will be approximately $1 - 2^{-10} \approx 0.999$. Typing additional loci will increase this probability even further.

5.2.1 Power of exclusion

The suitability of a set of genetic markers for paternity testing can be quantified by the **power of exclusion**, which is the expected value of the exclusion probability for a random man when the mother and child are sampled at random from the population. If we consider a single autosomal locus and assume that the population is at Hardy-Weinberg equilibrium, then this can be calculated with the help of the law of total probability:

$$\begin{aligned}
PE &= \sum_{M,C} \mathbb{P}(M, C) \cdot EP(M, C) \\
&= \sum_{i} \mathbb{P}(M = C = A_iA_i) \cdot (1 - p_i)^2 + \sum_{i \neq j} \mathbb{P}(M = A_iA_j, C = A_iA_i) \cdot (1 - p_i)^2 + \\
&\quad \sum_{i \neq j} \mathbb{P}(M = A_jA_j, C = A_iA_j) \cdot (1 - p_i)^2 + \sum_{i \neq j \neq k} \mathbb{P}(M = A_jA_k, C = A_iA_j) \cdot (1 - p_i)^2 + \\
&\quad \sum_{i<j} \mathbb{P}(M = C = A_iA_j) \cdot (1 - p_i - p_j)^2 \\
&= \left(\sum_i p_i^2 \cdot \frac{1}{2} p_i + \sum_{i<j} 2p_ip_j \cdot \frac{1}{2} p_i + \sum_{i<j} p_j^2 \frac{1}{2} p_i + \sum_{i<j<k} 2p_jp_k \frac{1}{2} p_i \right) \cdot (1 - p_i)^2 +
\end{aligned}$$

$$\begin{aligned} & \sum_{i<j} 2p_i p_j \frac{1}{2}(p_i + p_j) \cdot (1 - p_i - p_j)^2 \\ = & \sum_i p_i(1 - p_i + p_i^2)(1 - p_i)^2 + \sum_{i<j} p_i p_j (p_i + p_j) \cdot (1 - p_i - p_j)^2. \end{aligned}$$

Here we have used the product rule

$$\mathbb{P}(M, C) = \mathbb{P}(M) \cdot \mathbb{P}(C|M)$$

to evaluate the joint probability of the genotypes of the mother and her offspring. Of course, the exclusion probability for a specific pair of genotypes can be either substantially less than or substantially greater than PE depending on the rarity of the alleles involved. Furthermore, if individuals are genotyped at multiple unlinked loci, then the overall power of exclusion is given by the formula

$$PE_{ml} = 1 - \prod_{l=1}^{L} (1 - PE_l),$$

where PE_l is the power of exclusion at the l'th locus.

5.2.2 Power of exclusion of paternal relatives

We can also ask how much power a set of markers has to rule out paternity for male relatives of the true father. Let R be a male relative of the true father F and let $k_0, 2k_1, k_2$ be the relatedness coefficients of R and F. If we let Z be the number of alleles that are identical by descent (ibd) in R and C, then under the assumption that the mother is unrelated to F and R we have

$$\begin{aligned} \mathbb{P}(Z = 0) &= k_0 + \frac{1}{2} \cdot (2k_1) = k_0 + k_1 \\ \mathbb{P}(Z = 1) &= \frac{1}{2} \cdot (2k_1) + k_2 = k_1 + k_2 \\ \mathbb{P}(Z = 2) &= 0. \end{aligned}$$

If R and C do share an allele that is ibd, then paternity cannot be excluded for R. However, if R and C share no alleles that are ibd, then the genotypes of R and C and that locus are independent of one another and so the probability of excluding R as the father is the same as the probability of excluding a randomly sampled man from the population. (Here we are ignoring population structure.) It follows that the power of exclusion for a relative R is equal to

$$PE_R = \mathbb{P}(Z = 0) \cdot PE = (k_0 + k_1) \cdot PE.$$

This shows that the power of exclusion decreases as the relatedness between the alleged father and the true father increases. For example, if R and F are full brothers, then $k_0 = k_1 = 1/4$ and so $PE_R = \frac{1}{2} \cdot PE$.

5.3 Mutation

Recall that STRs have enjoyed widespread use as forensic markers in part because these loci are highly mutable and therefore highly polymorphic. However, high mutation rates can be problematic for paternity testing since transmission of a mutated allele from the father to the child could lead to a false exclusion. For this reason, the American Association of Blood Banks recommends that paternity only be excluded if there are mismatches between the alleged father and the child at at least two loci. When a single mismatch is detected, the paternity index is

replaced by the **average mutation paternity index**, which is equal to the following likelihood ratio

$$\begin{aligned} AMPI &= \frac{\mathbb{P}(C, M, AF \nsim C|H_p)}{\mathbb{P}(C, M, AF \nsim C|H_d)} \\ &= \frac{\mathbb{P}(AF \nsim C|C, M, H_p)}{\mathbb{P}(AF \nsim C|C, M, H_d)} \cdot \frac{\mathbb{P}(C, M|H_p)}{\mathbb{P}(C, M|H_d)} \\ &= \frac{\mathbb{P}(AF \nsim C|C, M, H_p)}{\mathbb{P}(AF \nsim C|C, M, H_d)}, \end{aligned}$$

where H_p is the hypothesis that AF is the father of C, H_d is the hypothesis that the father is a random male from population $\mathcal{P}$, and the notation $AF \nsim C$ means that the genotypes of AF and C share no alleles in common. As before, we continue to assume that mating is random, i.e., the genotypes of the mother and father are independent, which justifies the cancellation of the terms in the second ratio on the second line. Also, under the hypothesis that AF is the father of C, $AF \nsim C$ will be true only if the allele transmitted from AF to C has mutated (ignoring typing errors) and so

$$\mathbb{P}(AF \nsim C|C, M, H_p) = \mu$$

where μ is the average mutation rate per generation at that locus. On the other hand, the probability in the denominator is just the exclusion probability for that locus since AF will be excluded as the father if and only if $AF \nsim C$. This shows that the AMPI is equal to

$$AMPI = \frac{\mu}{EP},$$

which will typically be of the same order as the mutation rate. The overall paternity index, taking into account the data at all loci, is calculated by multiplying the AMPI by the paternity indices at the other loci.

Notice that value of the $AMPI$ depends on the fact that the genotypes of the alleged father and the child are incompatible, but does not depend on the specific genotype of AF. In some cases, one may be able to calculate a more powerful statistic by considering the specific mutation required to convert a paternal allele to the child's allele. This usually requires a mathematical model of the mutation process at the locus such as the stepwise mutation model for an STR which postulates that the copy number can either increase or decrease by at most one repeat unit per generation. Under this model, paternity would be excluded if all of the alleles of the alleged father and of the child differ by at least two repeats.

Chapter 6

Kinship Testing

6.1 Kinship between two persons

Paternity is just one of the forms of kinship that can be inferred using genetic data. Suppose that the genotypes of two individuals, Y and Z, are known and that we wish to determine whether these two individuals are related in some specific way, e.g., whether Y and Z are siblings. As with paternity testing, a common approach is to calculate the likelihood ratio for the following pair of hypotheses,

$$\begin{aligned} &H_p : Y \text{ and } Z \text{ are related;} \\ &H_d : Y \text{ and } Z \text{ are random, unrelated members of } \mathcal{P}, \end{aligned}$$

where $\mathcal{P}$ is the population to which the two individuals belong and H_p specifies the particular relationship that is alleged to hold between Y and Z. In fact, if we ignore mutation and only consider the genetic data, then these hypotheses can be reformulated as

$$\begin{aligned} &H_p : (Y, Z) \sim (k_0, 2k_1, k_2); \\ &H_d : (Y, Z) \sim (1, 0, 0), \end{aligned}$$

where $(k_0, 2k_1, k_2)$ are the relatedness coefficients of the two individuals under the alleged relationship. (These hypothesis pairs are, in general, not equivalent if mutation is taken into account, since different kinds of biological relationships may have the same relatedness coefficients, but produce different genotype distributions under certain models of mutation. Likewise, non-genetic evidence may contain information that allows us to discriminate between relationships that have the same relatedness coefficients, e.g., half siblings vs. uncle-nephew). The likelihood ratio can then be written as

$$\begin{aligned} LR &= \frac{\mathbb{P}(Y, Z|H_p)}{\mathbb{P}(Y, Z|H_d)} \\ &= \frac{\mathbb{P}(Z|Y, H_p)\mathbb{P}(Y|H_p)}{\mathbb{P}(Z|Y, H_d)\mathbb{P}(Y|H_d)} \\ &= \frac{\mathbb{P}(Z|Y, H_p)}{\mathbb{P}(Z|Y, H_d)} \\ &= \frac{\mathbb{P}(Z|Y, H_p)}{\mathbb{P}(Z)}, \end{aligned} \tag{6.1}$$

since the genotypes of Z and Y are conditionally independent under H_d. The numerator of the likelihood ratio can be calculated by conditioning on the number of alleles I that are ibd between Y and Z and using the law of total probability. If we assume that the population is at

Hardy-Weinberg equilibrium, then

$$\begin{aligned}\mathbb{P}(Z|Y,H_p) &= \sum_{t=0}^{2}\mathbb{P}(Z|Y,H_p,I=t)\cdot\mathbb{P}(I=t|H_p)\\ &= \mathbb{P}(Z|Y,H_p,I=0)\cdot k_0+\mathbb{P}(Z|Y,H_p,I=1)\cdot 2k_1+\mathbb{P}(Z|Y,H_p,I=2)\cdot 2k_2\\ &= \mathbb{P}(Z)\cdot k_0+\mathbb{P}(Z|Y,H_p,I=1)\cdot 2k_1+1_{\{Y=Z\}}\cdot k_2,\end{aligned}$$

where

$$1_{\{Y=Z\}}=\begin{cases}1 & \text{if } Y=Z\\ 0 & \text{if } Y\neq Z,\end{cases}$$

is the indicator function for the event that Y and Z have the same genotypes. Substituting this expression into equation (6.1) gives

$$LR=k_0+\left(2k_1\cdot\mathbb{P}(Z|Y,H_p,I=1)+k_2\cdot 1_{\{Y=Z\}}\right)/\mathbb{P}(Z).$$

For example, if $Z=Y=A_iA_i$, then

$$\begin{aligned}LR &= k_0+\left(2k_1p_i+k_2\cdot 1\right)/p_i^2\\ &= k_0+2k_1/p_i+k_2/p_i^2,\end{aligned}$$

while if $Z=A_iA_j$ and $Y=A_iA_k$, where $j\neq k$, then

$$\begin{aligned}LR &= k_0+\left(2k_1\cdot\frac{1}{2}p_j+k_2\cdot 0\right)/2p_ip_j\\ &= k_0+k_1/(2p_i).\end{aligned}$$

Furthermore, if Y and Z share no alleles in common, then $LR=k_0$. Table 5.2 of Fung & Hu (2008) provides similar expressions for the other possible combinations of genotypes. Likelihood ratios for multiple unlinked loci can then be calculated by multiplying the appropriate single locus likelihood ratios; as in the past two chapters, this is valid as long as the loci are approximately at linkage equilibrium.

6.1.1 Pedigrees

Suppose that a family of eight individuals is described by the pedigree illustrated in the figure below (Figure 5.1 of Fung & Hu (2008)), where the sold lines indicate known relationships and the dashed lines correspond to suspected relationships. The question in this case is whether persons 7 and 8 are half sibs or first cousins, which leads to the following pair of hypotheses:

$$\begin{aligned}&H_{p1}:(Y,Z)\sim(0.5,0.5,0);\\ &H_{p2}:(Y,Z)\sim(0.75,0.25,0),\end{aligned}$$

where Y is the genotype of person 7, Z is the genotype of person 8, H_{p1} alleges that 7 and 8 are full siblings, and H_{p2} alleges that they are first cousins. Although the likelihood ratio LR for H_{p1} vs. H_{p2} could be evaluated directly, it can also be obtained from the following identity

$$LR=LR_1/LR_2,$$

where LR_1 is the likelihood ratio for the hypothesis pair

$$\begin{aligned}&H_{p1}:(Y,Z)\sim(0.5,0.5,0);\\ &H_d:(Y,Z)\sim(1,0,0),\end{aligned}$$

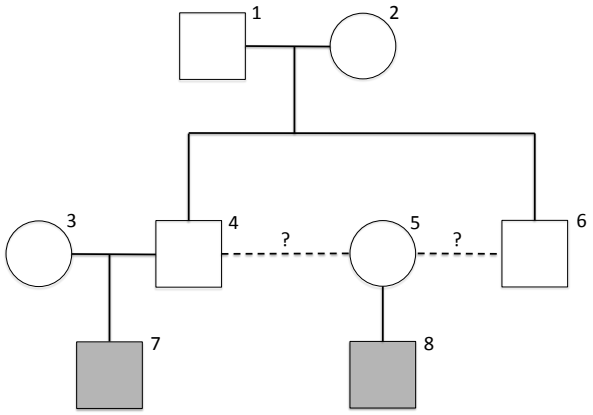

Figure 6.1: Pedigree with known and suspected relationships.

and $LR2$ is the likelihood ratio for the hypothesis pair

$$H_{p2} : (Y, Z) \sim (0.75, 0.25, 0);$$
$$H_d : (Y, Z) \sim (1, 0, 0).$$

The only merit (as far as I can tell) for solving the problem in this manner is that one can look up the relevant formulas for $LR1$ and $LR2$ in Table 5.2 of Fung & Hu (2008). For example, if $Y = Z = A_iA_i$, then

$$\begin{aligned} LR1 &= \frac{1}{2} + \frac{1}{2p_i} \\ LR2 &= \frac{3}{4} + \frac{1}{4p_i}, \end{aligned}$$

giving

$$LR = \frac{2 + 2p_i}{1 + 3p_i},$$

which is a decreasing function of p_i bounded between 1 and 2, i.e., shared homozygous genotypes provide stronger support for the closer relationship.

6.1.2 Subdivided populations

If the population is subdivided and persons Y and Z belong to the same subpopulation, then the appropriate hypothesis pair is

$H_p : Y$ and Z are related;
$H_d : Y$ and Z are random, unrelated members of the same subpopulation,

with likelihood ratio

$$LR = \frac{\mathbb{P}(Z|Y, H_p)}{\mathbb{P}(Z|Y, H_d)}.$$

In this case, the conditional probability appearing in the denominator does not simplify since the genotype of Z is not independent of that Y if only the global allele frequencies are known. Nonetheless, if we assume that the island model introduced in Section 3.2 applies, then the

conditional probabilities appearing in the numerator and denominator can be calculated by repeatedly applying the conditional sampling distribution formula (3.14). For example, suppose that $Y = Z = A_iA_i$. Under H_d, the two individuals share no alleles that are identical by descent and so

$$\begin{aligned}\mathbb{P}(Z = A_iA_i|Y = A_iA_i, H_d) &= \mathbb{P}(A_i|A_iA_i) \cdot \mathbb{P}(A_i|A_iA_iA_i) \\ &= \frac{2\theta + (1-\theta)p_i}{1+\theta} \frac{3\theta + (1-\theta)p_i}{1+2\theta}.\end{aligned}$$

As above, the conditional probability in the numerator can be evaluated by conditioning on the number of alleles I that are identical by descent between Y and Z:

$$\mathbb{P}(Z|Y, H_p) = \mathbb{P}(Z|Y, H_p, I = 0) \cdot k_0 + \mathbb{P}(Z|Y, H_p, I = 1) \cdot 2k_1 + \mathbb{P}(Z|Y, H_p, I = 2) \cdot 2k_2,$$

but now we must take this number into account when applying the conditional sampling formula. Continuing with our example and considering each of the three cases separately, we have

$$\begin{aligned}\mathbb{P}(Z = A_iA_i|Y = A_iA_i, H_p, I = 0) &= \mathbb{P}(Z = A_iA_i|Y = A_iA_i, H_d) \\ \mathbb{P}(Z = A_iA_i|Y = A_iA_i, H_p, I = 1) &= \mathbb{P}(A_i|A_iA_i) \\ &= \frac{2\theta + (1-\theta)p_i}{1+\theta} \\ \mathbb{P}(Z = A_iA_i|Y = A_iA_i, H_p, I = 2) &= 1.\end{aligned}$$

It follows that the likelihood ratio for this example is equal to

$$LR = k_0 + \frac{2k_1(1+2\theta)}{3\theta + (1-\theta)p_i} + \frac{k_2(1+\theta)(2+\theta)}{(2\theta + (1-\theta)p_i)(3\theta + (1-\theta)p_i)},$$

which in general is a decreasing function of θ, i.e., shared homozygous genotypes provide weaker evidence of consanguinity in a subdivided population than in a panmictic population. As before, we will have $LR = k_0$ whenever the two genotypes have no alleles in common. Explicit formulas for the other possible combinations of genotypes can be found in Table 5.5 of Fung & Hu (2008).

6.2 Kinship involving three persons

Suppose that the genotypes of three persons X, Y and Z are available and that we wish to determine the relationships between these individuals. Although this problem could be divided into three two-person kinship problems, a more powerful approach is to analyze all of the genetic data simultaneously. We will follow Fung & Hu (2008) in assuming that X and Z are unrelated and that all three individuals belong to the same population, which is at Hardy-Weinberg equilibrium. We will also write $(k_0^{XY}, 2k_1^{XY}, k_2^{XY})$ and $(k_0^{YZ}, 2k_1^{YZ}, k_2^{YZ})$ for the relatedness coefficients for X and Y and for Y and Z, respectively. Notice that if X and Z are unrelated, but each individual is related to Y, then we must have $k_2^{XY} = k_2^{YZ} = 0$, since otherwise the relatedness coefficient,

$$2k_1^{XZ} \geq k_2^{XY} \cdot 2k_1^{YZ} + 2k_1^{XY} \cdot k_2^{YZ} > 0,$$

would necessarily be positive.

To calculate the likelihood ratios for different hypotheses about the values of the two sets of relatedness coefficients, we need to be able to calculate the joint probabilities of the three genotypes X, Y and Z. Formulas covering the possible combinations of genotypes are given in Figure 5.10 in Fung & Hu (2008). Here we simply illustrate the derivations of these formulas for a few

special cases. Suppose first that $X = Y = Z = A_iA_i$. Then, by conditioning on the numbers of alleles that are bid between X and Y and between Y and Z, the law of total probability gives

$$\begin{aligned}
\mathbb{P}(X = Y = Z = A_iA_i) \;=\; & \mathbb{P}(X = Y = Z = A_iA_i | I^{XY} = 0, I^{YZ} = 0) \cdot k_0^{XY} k_0^{YZ} + \\
& \mathbb{P}(X = Y = Z = A_iA_i | I^{XY} = 1, I^{YZ} = 0) \cdot 2k_1^{XY} k_0^{YZ} + \\
& \mathbb{P}(X = Y = Z = A_iA_i | I^{XY} = 0, I^{YZ} = 1) \cdot 2k_0^{XY} k_1^{YZ} + \\
& \mathbb{P}(X = Y = Z = A_iA_i | I^{XY} = 1, I^{YZ} = 1) \cdot 4k_1^{XY} k_1^{YZ} \\
= \;& \mathbb{P}(Y = A_iA_i | X = Z = A_iA_i, I^{XY} = 0, I^{YZ} = 0) \cdot \mathbb{P}(X = Z = A_iA_i) \cdot k_0^{XY} k_0^{YZ} + \\
& \mathbb{P}(Y = A_iA_i | X = Z = A_iA_i, I^{XY} = 1, I^{YZ} = 0) \cdot \mathbb{P}(X = Z = A_iA_i) \cdot 2k_1^{XY} k_0^{YZ} + \\
& \mathbb{P}(Y = A_iA_i | X = Z = A_iA_i, I^{XY} = 0, I^{YZ} = 1) \cdot \mathbb{P}(X = Z = A_iA_i) \cdot 2k_0^{XY} k_1^{YZ} + \\
& \mathbb{P}(Y = A_iA_i | X = Z = A_iA_i, I^{XY} = 1, I^{YZ} = 1) \cdot \mathbb{P}(X = Z = A_iA_i) \cdot 4k_1^{XY} k_1^{YZ} \\
= \;& p_i^2 \cdot \left(p_i^2\right)^2 \cdot k_0^{XY} k_0^{YZ} + p_i \cdot \left(p_i^2\right)^2 \cdot 2k_1^{XY} k_0^{YZ} + p_i \cdot \left(p_i^2\right)^2 \cdot 2k_0^{XY} k_1^{YZ} + 1 \cdot \left(p_i^2\right)^2 \cdot 4k_1^{XY} k_1^{YZ} \\
= \;& \left\{ p_i^2 \cdot k_0^{XY} k_0^{YZ} + p_i \cdot 2(k_1^{XY} k_0^{YZ} + k_0^{XY} k_1^{YZ}) + 4k_1^{XY} k_1^{YZ} \right\} \cdot p_i^4.
\end{aligned}$$

Several comments are in order. First, in passing from the first to the second expression, we have used the fact that the genotypes of X and Z are independent of the numbers of alleles that these individuals share ibd with Y, e.g.,

$$\begin{aligned}
\mathbb{P}(X = Z = A_iA_i | I^{XY} = 0, I^{YZ} = 0) \;&=\; \mathbb{P}(X = Z = A_iA_i) \\
&=\; \mathbb{P}(X = A_iA_i)\mathbb{P}(Z = A_iA_i) = p_i^4,
\end{aligned}$$

where the second line follows from the fact that X and Z are assumed to be unrelated members of a population at HWE. Furthermore, because X and Z share no alleles that are ibd, it follows that if X and Y share one allele that is ibd and Z and Y share one allele that is ibd, then these must be different alleles. This implies that

$$\mathbb{P}(Y = A_iA_i | X = Z = A_iA_i, I^{XY} = 1, I^{YZ} = 1) = 1,$$

since each of the alleles present in Y is ibd with an allele in X or an allele in Z. Similarly, if $X = A_iA_j$, $Y = A_iA_i$, and $Z = A_kA_l$, where $i \neq j \neq k \neq l$, then we must have $I^{YZ} = 0$, since Y and Z have no alleles that are identical by state. It follows that

$$\begin{aligned}
\mathbb{P}(X = A_iA_j, Y = A_iA_i, Z = A_kA_l) \;=\; & \\
& \mathbb{P}(X = A_iA_j, Y = A_iA_i, Z = A_kA_l | I^{XY} = 0, I^{YZ} = 0) \cdot k_0^{XY} k_0^{YZ} + \\
& \mathbb{P}(X = A_iA_j, Y = A_iA_i, Z = A_kA_l | I^{XY} = 1, I^{YZ} = 0) \cdot 2k_1^{XY} k_0^{YZ} \\
=\; & \mathbb{P}(Y = A_iA_i | X = A_iA_j, I^{XY} = 0) \cdot \mathbb{P}(X = A_iA_j) \cdot \mathbb{P}(Z = A_kA_l) \cdot k_0^{XY} k_0^{YZ} + \\
& \mathbb{P}(Y = A_iA_i | X = A_iA_j, I^{XY} = 1) \cdot \mathbb{P}(X = A_iA_j) \cdot \mathbb{P}(Z = A_kA_l) \cdot 2k_1^{XY} k_0^{YZ} \\
=\; & p_i^2 \cdot 2p_ip_j \cdot 2p_kp_l \cdot k_0^{XY} k_0^{YZ} + \frac{1}{2} p_i \cdot 2p_ip_j \cdot 2p_kp_l \cdot 2k_1^{XY} k_0^{YZ} \\
=\; & \left\{ p_i^2 \cdot k_0^{XY} + p_i \cdot 2k_1^{XY} \right\} \cdot k_0^{YZ} \cdot 4p_ip_jp_kp_l.
\end{aligned}$$

These formulas can be used to calculate likelihood ratios as follows. If no non-genetic evidence is available and mutation is neglected, then the pair of hypotheses to be compared can be written as

$$\begin{aligned}
& H_{p1} : (X, Y) \sim (k_0^{XY}, 2k_1^{XY}, k_2^{XY}), (Y, Z) \sim (k_0^{YZ}, 2k_1^{YZ}, k_2^{YZ}); \\
& H_{p2} : (X, Y), (Y, Z) \sim (1, 0, 0),
\end{aligned}$$

and the likelihood ratio is equal to

$$LR \;=\; \frac{\mathbb{P}(X, Y, Z | H_p)}{\mathbb{P}(X, Y, Z | H_d)} \;=\; \frac{\mathbb{P}(X, Y, Z | H_p)}{\mathbb{P}(X)\mathbb{P}(Y)\mathbb{P}(Z)}$$

where we have used the assumption that the population is at HWE to simplify the denominator.

Chapter 7

DNA Mixtures

7.1 Two contributors

When a biological sample contains DNA from more than one contributor, it may not be possible to assign diploid genotypes to individual contributors and then likelihood ratio calculations must take into account the various possible combinations of genotypes that are consistent with the mixture. In this section we will examine three common scenarios that involve mixtures of DNA from two contributors. Throughout we will consider a single autosomal locus segregating alleles $A_1, \cdots, A_n$ at frequencies $p_1, \cdots, p_n$ and we will assume that the population is at Hardy-Weinberg equilibrium.

7.1.1 One victim and one suspect

Suppose that the mixture M contains at most four distinct alleles and that the two competing hypotheses are

$H_p : M$ is a mixture of DNA from the victim and the suspect;
$H_d : M$ is a mixture of DNA from the victim and a random member of the population.

We will assume that the genotypes of both the victim V and the suspect S are known and are denoted by K. Then the likelihood ratio is equal to

$$\begin{aligned} LR &= \frac{\mathbb{P}(M, K|H_p)}{\mathbb{P}(M, K|H_d)} \\ &= \frac{\mathbb{P}(M|K, H_p)}{\mathbb{P}(M|K, H_d)} \cdot \frac{\mathbb{P}(K|H_p)}{\mathbb{P}(K|H_d)} \\ &= \frac{\mathbb{P}(M|K, H_p)}{\mathbb{P}(M|K, H_d)}, \end{aligned}$$

since the genotypes specified in K are independent of the two hypotheses. To illustrate how these probabilities are calculated, suppose that $M = \{A_i, A_j, A_k\}$ is a mixture of three alleles and that the victim's genotype is A_iA_j and that the suspect's genotype is A_iA_k. Then

$$\mathbb{P}(M|K, H_p) = 1$$

since K contains the same three alleles as M, i.e., the genotypes of the victim and suspect suffice to explain the genetic composition of the mixture. To calculate the conditional probability in the denominator, let X be the genotype of the unknown contributor posited by H_d and notice

that we must have $X = A_iA_k$, $X = A_jA_k$ or $X = A_kA_k$ for V and X to explain the mixture. Provided that X and V are unrelated and that the population is at HWE, we have

$$\begin{aligned}\mathbb{P}(M|K,H_d) &= \mathbb{P}(X = A_iA_k|K,H_d) + \mathbb{P}(X = A_jA_k|K,H_d) + \mathbb{P}(X = A_kA_k|K,H_d)\\ &= \mathbb{P}(X = A_iA_k) + \mathbb{P}(X = A_jA_k) + \mathbb{P}(X = A_kA_k)\\ &= 2p_ip_k + 2p_jp_k + p_k^2,\end{aligned}$$

from which it follows that

$$LR = \frac{1}{p_k(2p_i + 2p_j + p_k)}.$$

Notice that this quantity is smaller than any of the three likelihood ratios that would have been calculated if the sample contained DNA only from the suspect, e.g., if the sample is typed and found to contain the alleles A_iA_k, then the match probability for a random member of the population is $2p_ip_k$ and so the likelihood ratio would be $1/2p_ip_k$, which is larger than that calculated for the mixture. There is a loss of power in this case because the mixture does not uniquely determine the genotype of the other contributor. On the other hand, if the victim's genotype were A_jA_j, then the genotype of the other contributor would have to be A_iA_k. In this case we still have $\mathbb{P}(M|K,H_p) = 1$, while $\mathbb{P}(M|K,H_d) = \mathbb{P}(X = A_iA_k) = 2p_ip_k$ and so the likelihood ratio is

$$LR = \frac{1}{2p_ip_k},$$

which is the same value that was obtained above for data from a single contributor. This occurs because the victim's profile explains the presence of the allele A_j in the sample and allows us to unambiguously assign a diploid genotype to the second contributor.

7.1.2 One suspect and one unknown

If a suspect has been identified but the second contributor to a mixed sample is unknown, then the prosecution and defense hypotheses may be

H_p : M is a mixture of DNA from the suspect and a random member of the population;
H_d : M is a mixture of DNA from two random members of the population.

Let K denote the known genotype of the suspect and let X denote the genotypes of the unknown contributors. Then X contains either one or two genotypes depending on whether we accept H_p or H_d. As before, suppose that $M = \{A_i, A_j, A_k\}$ and that the suspect's genotype is A_iA_j. Under H_p, we must have $X = A_iA_k$, $X = A_jA_k$ or $X = A_kA_k$ for S and X to explain the mixture and so

$$\mathbb{P}(M|K,H_p) = 2p_ip_k + 2p_jp_k + p_k^2.$$

In contrast, under H_d, X contains alleles from two contributors and these must account for the three alleles observed in the mixture. Consultation of Table 6.1 in Fung & Hu (2008) shows that there are twelve possible combinations of genotypes that are consistent with the mixture and that the sum of the probabilities of these cases is equal to

$$\mathbb{P}(M|K,H_d) = \mathbb{P}(M|H_d) = 12p_ip_jp_k\left(p_i + p_j + p_k\right).$$

Thus the likelihood ratio in this case is equal to

$$LR = \frac{2p_i + 2p_j + p_k}{12p_ip_j\left(p_i + p_j + p_k\right)} = \frac{1}{2p_ip_j} \cdot \left(\frac{2p_i + 2p_j + p_k}{6p_i + 6p_j + 6p_k}\right),$$

which is less than the likelihood ratio $1/2p_ip_j$ for the case without a second contributor.

7.1.3 Two suspects

If there are two unrelated suspects, S_1 and S_2, then the prosecution and defense hypotheses could be

$$H_p : M \text{ is a mixture of DNA from the two suspects;}$$
$$H_d : M \text{ is a mixture of DNA from two random members of the population.}$$

In this case we will have

$$\mathbb{P}(M|K, H_p) = 1$$

as long as the genotypes K of the two suspects are consistent with the mixture. Furthermore, if the population is at HWE and the suspects are unrelated to the true contributors, then M is independent of K under H_d and so

$$\mathbb{P}(M|K, H_d) = \mathbb{P}(M|H_d).$$

For example, if $M = \{A_i, A_j, A_k\}$, then

$$\mathbb{P}(M|K, H_d) = 12 p_i p_j p_k \left(p_i + p_j + p_k\right),$$

as calculated above, and so the likelihood ratio is just the reciprocal of this quantity

$$LR = \frac{1}{12 p_i p_j p_k \left(p_i + p_j + p_k\right)}.$$

7.2 Multiple contributors

As the number of contributors to a sample increases, the enumeration of all possible combinations of genotypes that are consistent with the available data will become too cumbersome to do by hand. For example, suppose that the sample contains five alleles $M = \{A_1, A_2, A_3, A_4, A_5\}$ and that the genotype of only one of the contributors is known, say $S = A_1A_2$. In this case, there must be at least two more contributors to account for the remaining three alleles. If we posit that there are exactly two more contributors, then there are 24 possible combinations of genotypes that will explain the mixture. Fortunately, there is a general recursive formula that will allow us to calculate the conditional probabilities for arbitrarily complicated mixtures without having to list all of the possibilities.

We first establish some notation. As above, let M denote the alleles in the mixture and let K denote the alleles of all the contributors that have been typed. We will also let x and X denote the number and alleles of the unknown contributors. Finally, let U be the set of alleles that are present in the M but not in K, i.e., $U = M \backslash K$. In general, we will be interested in calculating probabilities of the form

$$\mathbb{P}(M|K, H) = \mathbb{P}(U \subset X \subset M|K),$$

where H is a hypothesis that stipulates the known contributors to the mixture as well as the number of unknown contributors. Since

$$X \subset M = (U \subset X \subset M) \cup \left(\bigcup_{A_i \in U} (X \subset M \backslash \{A_i\}) \right)$$

is a disjoint union, it follows that

$$\mathbb{P}(X \subset M|K) = \mathbb{P}(U \subset X \subset M|K) + \mathbb{P}\left(\bigcup_{A_i \in U} (X \subset M \backslash \{A_i\}) | K \right),$$

or equivalently

$$\mathbb{P}(U \subset X \subset M|K) = \mathbb{P}(X \subset M|K) - \mathbb{P}\left(\bigcup_{A_i \in U} (X \subset M\backslash\{A_i\})|K\right). \tag{7.1}$$

Next, let

$$B_i = (X \subset M\backslash\{A_i\})$$

and observe that

$$\begin{aligned} B_i \cap B_j &= (X \subset M\backslash\{A_i, A_j\}) \\ B_i \cap B_j \cap B_k &= (X \subset M\backslash\{A_i, A_j, A_k\}) \\ &\cdots \end{aligned}$$

We can then use the inclusion-exclusion identity for probabilities to write

$$\begin{aligned} \mathbb{P}\left(\bigcup_{A_i \in U} (X \subset M\backslash\{A_i\})|K\right) &= \mathbb{P}\left(\bigcup_{A_i \in U} B_i|K\right) \\ &= \sum_{A_i \in U} \mathbb{P}(B_i|K) - \sum_{A_i, A_j \in U} \mathbb{P}(B_i \cap B_j|K) + \sum_{A_i, A_j, A_k \in U} \mathbb{P}(B_i \cap B_j \cap B_k|K) - \cdots \\ &= \sum_{A_i \in U} W(M\backslash\{A_i\}) - \sum_{A_i, A_j \in U} W(M\backslash\{A_i, A_j\}) + \sum_{A_i, A_j, A_k \in U} W(M\backslash\{A_i, A_j, A_k\}) \\ &\quad - \cdots + (-1)^{|U|+1} W(M\backslash U), \end{aligned}$$

where $|U|$ denotes the number of elements in U and

$$W(D) \equiv \mathbb{P}(X \subset D|K)$$

is defined for any subset D of M satisfying $M\backslash U \subset D \subset M$. Substituting this result into equation (7.1) shows that

$$\begin{aligned} \mathbb{P}(M|K, H) &= \sum_{M\backslash U \subset D \subset M} (-1)^{|M\backslash D|} W(D) \\ &= W(M) - \sum_{A_i \in U} W(M\backslash\{A_i\}) + \sum_{A_i, A_j \in U} W(M\backslash\{A_i, A_j\}) - \\ &\quad \sum_{A_i, A_j, A_k \in U} W(M\backslash\{A_i, A_j, A_k\}) + \cdots + (-1)^{|U|} W(M\backslash U), \end{aligned} \tag{7.2}$$

which is equation (6.7) in Fung & Hu (2008). Although this recursion may appear to be as complicated as the original problem, it can easily be implemented computationally and it is valid even when we do not assume that the population is at HWE.

7.2.1 Unknown contributors under Hardy-Weinberg equilibrium

As above, let x denote the number of unknown contributors to M posited by the hypothesis H and write $X_1, \cdots, X_{2x}$ for the $2x$ alleles carried by these individuals at an autosomal locus. If the population is at HWE and the unknown contributors are unrelated to one another and to the known contributors, then these alleles will be both mutually independent and independent of K and so

$$W(D) = \mathbb{P}(X \subset D|K)$$

$$\begin{aligned} &= \mathbb{P}(X_1, \cdots, X_{2x} \in D) \\ &= \prod_{i=1}^{2x} \mathbb{P}(X_i \in D) \\ &= \left(\sum_{A_i \in D} p_i \right)^{2x}. \end{aligned}$$

For example, if the hypothesis H stipulates that $M = \{A_1, A_2, A_3\}$, $K = \emptyset$, and $x = 2$, then $U = M$ and equation (7.2) gives

$$\begin{aligned} \mathbb{P}(M|K,H) &= W(M) - W(A_1,A_2) - W(A_1,A_3) - W(A_2,A_3) + \\ &\quad W(A_1) + W(A_2) + W(A_3) - W(\emptyset) \\ &= (p_1+p_2+p_3)^4 - (p_1+p_2)^4 - (p_1+p_3)^4 - (p_2+p_3)^4 + \\ &\quad p_1^4 + p_2^4 + p_3^4 \\ &= 12 p_1 p_2 p_3 (p_1+p_2+p_3). \end{aligned}$$

7.2.2 Unknown contributors from different ethnic groups

Although most genetic variation is shared across different human populations, there are some differences in the genetic composition of different ethnic groups that are large enough to make it necessary to take ethnicity into account when interpreting DNA mixtures. This is particularly important when dealing with rare alleles, since demographic processes such as genetic drift and founder effects can lead to large relative differences in the frequencies of these alleles in different populations. In turn, large relative differences in allele frequencies can give rise to large absolute differences in likelihood ratios depending on the assumptions that are made concerning the ethnicity of the contributors to the DNA mixture. These considerations are complicated by the fact that cultural and societal assumptions concerning ethnicity do not neatly align with the genetic structure of human populations and also by the fact that individuals may have biological ancestry that includes multiple ethnic groups.

We first consider a specific example and then derive a general identity that can be used in combination with equation (7.1) to calculate the probabilities of mixtures under arbitrarily complicated hypotheses that specify the ethnicities of the contributors. For our example, suppose that a mixture is found to contain four alleles $M = \{A_1, A_2, A_3, A_4\}$ and that two suspects with profiles $S_1 = A_1A_2$ and $S_2 = A_3A_4$ have been identified. In addition, suppose that the prosecution and defense have put forward the following two competing explanations of the mixture

H_p : M is a mixture of DNA from S_1 and S_2;
H_d : M is a mixture of DNA from two unknown persons: X_1 from ethnic group a and X_2 from ethnic group b.

Under the prosecution's hypothesis we have $\mathbb{P}(M|K,H_p) = 1$, where $K = \{S_1, S_2\}$ consists of the known profiles of the two suspects. To evaluate the probability under the defense's hypothesis, we need to know the frequencies of the alleles in the two ethnic groups. Suppose that the frequency of A_i is p_{ai} in ethnic group a and p_{bi} in ethnic group b. Since there are four alleles and only two contributors, each contributor must be heterozygous for a different pair of alleles in M and so there are six possibilities: $(A_1A_2, A_3A_4), (A_1A_3, A_2A_4), (A_1A_4, A_2A_3), \cdots$, where the first genotype in the ordered pair belongs to X_1 and the second one belongs to X_2. Summing over the probabilities of each of these outcomes gives

$$\mathbb{P}(M|K,H_d) = 4p_{a1}p_{a2}p_{b3}p_{b4} + 4p_{a1}p_{a3}p_{b2}p_{b4} + 4p_{a1}p_{a4}p_{b2}p_{b3}$$

$$+4p_{a2}p_{a3}p_{b1}p_{b4} + 4p_{a2}p_{a4}p_{b1}p_{b3} + 4p_{a3}p_{a4}p_{b1}p_{b2}.$$

It is clear from the example that exhaustive enumeration is not feasible when either the mixture or the hypothesis is more complicated. Fortunately, equation (7.1) still holds when there are genetic differences between ethnic groups provided that we use a suitable expression for the probabilities $W(D) = \mathbb{P}(X \in D|K)$. Here we will consider hypotheses H that stipulate the ethnicities of the unknown contributors. Let $G = \{a, b, \cdots\}$ be a set of indices that correspond to different ethnic groups (e.g., ethnic group a, ethnic group b, etc.) and for each $g \in G$ let p_{gi} denote the frequency of allele A_i in group g. We will write x_g for the number of unknown contributors from ethnic group g that is posited by H. If we assume that Hardy-Weinberg equilibrium holds separately in each ethnic group and also that the contributors are unrelated to one another, then the alleles contributed by these individuals will be conditionally independent of one another given the ethnicities of the contributors and so

$$W(D) = \mathbb{P}(X \in D|K) = \prod_{g \in G} \left(\sum_{A_i \in D} p_{gi} \right)^{2x_g}.$$

For example, suppose that $M = \{A_1, A_2, A_3\}$ and that we wish to test the hypothesis that there are two unknown contributors that come from different ethnic groups, say group a and group b. Under this hypothesis, we have $U = M$ and $x_a = x_b = 1$ and so

$$\begin{aligned}
\mathbb{P}(M|K, H_d) &= W(M) - W(M\backslash\{A_1\}) - W(M\backslash\{A_2\}) - W(M\backslash\{A_3\}) \\
&\quad + W(M\backslash\{A_1, A_2\}) + W(M\backslash\{A_1, A_3\}) + W(M\backslash\{A_2, A_3\}) \\
&= (p_{a1} + p_{a2} + p_{a3})^2(p_{b1} + p_{b2} + p_{b3})^2 - (p_{a2} + p_{a3})^2(p_{b2} + p_{b3})^2 \\
&\quad - (p_{a1} + p_{a3})^2(p_{b1} + p_{b3})^2 - (p_{a1} + p_{a2})^2(p_{b1} + p_{b2})^2 \\
&\quad + p_{a3}^2 p_{b3}^2 + p_{a2}^2 p_{b2}^2 + p_{a1}^2 p_{b1}^2.
\end{aligned}$$

7.2.3 Unknown contributors from a subdivided population

If the population is subdivided and there are multiple unknown contributors from the same unknown subpopulation, the genotypes of these contributors will usually not be mutually independent. Once again, we will start with a concrete example and then derive a general expression for the probability $W(D)$ that can be incorporated into equation (7.1).

Let us revisit the example from section 7.1.1 where we had a mixture $M = \{A_1, A_2, A_3\}$ (say) as well as a known victim $V = A_1A_2$ and known suspect $S = A_1A_3$. However, we will now modify the defense hypothesis H_d so that the unknown contributor X is a random member of the same subpopulation as the victim and suspect. As before, under H_d, X must have one of the following three genotypes: A_1A_3, A_2A_3 or A_3A_3. However, because X, V and S belong to the same subpopulation, the known genotypes of V and S are not independent of the unknown genotype of X. Fortunately, the sampling distribution derived for the island model in section 3.2.3 can be used to account for this dependence provided that we have an estimate of the inbreeding coefficient θ for the population. Using the same short-hand notation for samples of alleles, we have

$$\begin{aligned}
\mathbb{P}(M|K, H_d) &= \mathbb{P}(X = A_1A_3|V = A_1A_2, S = A_1A_3) + \\
&\quad \mathbb{P}(X = A_2A_3|V = A_1A_2, S = A_1A_3) + \\
&\quad \mathbb{P}(X = A_3A_3|V = A_1A_2, S = A_1A_3) \\
&= 2\mathbb{P}(A_1, A_3|A_1, A_1, A_2, A_3) + 2\mathbb{P}(A_1, A_3|A_1, A_1, A_2, A_3) +
\end{aligned}$$

$$
\begin{aligned}
& \mathbb{P}(A_3, A_3|A_1, A_1, A_2, A_3) \\
= \; & 2\mathbb{P}(A_1|A_1, A_1, A_2, A_3) \cdot \mathbb{P}(A_3|A_1, A_1, A_1, A_2, A_3) + \\
& 2\mathbb{P}(A_2|A_1, A_1, A_2, A_3) \cdot \mathbb{P}(A_3|A_1, A_1, A_2, A_2, A_3) + \\
& \mathbb{P}(A_3|A_1, A_1, A_2, A_3) \cdot \mathbb{P}(A_3|A_1, A_1, A_2, A_3, A_3) \\
= \; & 2\left[\frac{2\theta + (1-\theta)p_1}{1+3\theta}\right]\left[\frac{\theta + (1-\theta)p_3}{1+4\theta}\right] + 2\left[\frac{\theta + (1-\theta)p_2}{1+3\theta}\right]\left[\frac{\theta + (1-\theta)p_3}{1+4\theta}\right] + \\
& \left[\frac{\theta + (1-\theta)p_3}{1+3\theta}\right]\left[\frac{2\theta + (1-\theta)p_3}{1+4\theta}\right].
\end{aligned}
$$

For the general case, we need to be able to calculate the probabilities $W(D) = \mathbb{P}(X \subset D|K)$. Fung & Hu (2008) describe one approach based on combinatorial analysis of expressions of the form

$$r^{(m)}(k, \theta) = \prod_{i=0}^{m-1} \Big((k+i)\theta + (1-\theta)r\Big),$$

but the final result can also be obtained directly from the island model. Recall the assumptions of that model: we have a single population divided into a large number of demes that regularly exchange migrants and we are interested in the distribution of allele frequencies within demes conditional on the global allele frequencies, which are taken to be $p_1, \cdots, p_n$. Furthermore, since we do not take mutation or selection into account, the alleles themselves are completely exchangeable in this model, provided that we apply the same relabeling to each subpopulation. In particular, we can merge sets of alleles into single allelic states and the sampling formula from section 3.2.3 will still remain valid. For example, suppose that the population is segregating four alleles $A_1, \cdots, A_4$ at frequencies $p_1, \cdots, p_4$, but that we do not wish to distinguish between individuals carrying the A_1 or A_2 allele. Then we can simply define a new allelic class, say A_{12}, which contains all of these individuals and which has population frequency p_{12}, and the island model will still apply when we consider the three allelic classes A_{12}, A_3, and A_4.

To see how this observation can help us to calculate $W(D)$, let $c_i, 1 \leq i \leq n$, denote the number of A_i alleles carried by the individuals with known genotypes and let $c = c_1 + \cdots + c_n$ be the total number of chromosomes carried by these individuals. We will define a new allelic class, A_D, by merging all of the alleles belonging to D and let

$$
\begin{aligned}
p_D &= \sum_{A_l \in D} p_l \\
c_D &= \sum_{A_l \in D} c_l
\end{aligned}
$$

denote the frequency of A_D in the population and the number of A_D alleles carried by the known individuals, respectively. Since the sampling formula still applies, it follows that

$$
\begin{aligned}
W(D) &= \mathbb{P}(2x \text{ additional } A_D \text{ alleles are sampled}|\ c_D \text{ were sampled out of } c) \\
&= \prod_{i=0}^{2x-1} \mathbb{P}(\text{an additional } A_D \text{ is sampled}|c_D + i \text{ copies of } A_D \text{ among } c+i \text{ alleles}) \\
&= \prod_{i=0}^{2x-1} \frac{(c_D + i)\theta + (1-\theta)p_D}{1 + (c+i-1)\theta} \\
&= \frac{p_D^{(2x)}(c_D, \theta)}{1^{(2x)}(c, \theta)}.
\end{aligned}
$$

For example, to apply this method to the Simpson case, we take $c_1 = 2$ and $c_2 = c_3 = 1$,

obtaining

$$\begin{aligned}
\mathbb{P}(M|K,H_d) &= W(M) - W(M\backslash\{A_3\}) \\
&= W(A_1,A_2,A_3) - W(A_1,A_2) \\
&= \frac{4\theta+(1-\theta)(p_1+p_2+p_3}{1+3\theta}\cdot\frac{5\theta+(1-\theta)(p_1+p_2+p_3}{1+4\theta} - \\
&\quad \frac{3\theta+(1-\theta)(p_1+p_2)}{1+3\theta}\cdot\frac{4\theta+(1-\theta)(p_1+p_2)}{1+4\theta},
\end{aligned}$$

which can be further simplified with the help of a computer algebra package if desired.

Chapter 8

Statistical Phylogenetics

8.1 Introduction

Our focus in this section will be on the analysis of DNA sequence data and some of its forensic applications, especially in the growing field of **forensic microbiology**. Much of the work in this area has focused on the transmission of infectious diseases, either in a healthcare setting or through criminal activity including bioterrorism. One of the key challenges for investigators is to be able to identify the source of a new infection, which can sometimes be done by comparing the genomes of pathogens isolated from victims with the genomes of pathogens sampled from the environment. Some of the more prominent cases have included:

- HIV-1 transmission by a Florida dentist (Ou et al., 1992);
- transmission of HIV-1 during a sexual assault (Sweden: Albert et al., 1994);
- deliberate transmission of HIV-1 (Louisiana: Metzker et al., 2002);
- anthrax attacks: 1993 (Kameido, Japan), 2001 Amerithrax cases (US).

There are a number of important differences between the aims, the methods, and the data used in human forensic DNA profiling and forensic microbiology. In particular,

- Aims: Genetic profiling seeks to identify the individual contributors of a sample of DNA, whereas forensic microbiology is usually concerned with more distant relationships between individuals and populations.
- Data: Genetic profiling still largely relies on STR markers, whereas forensic microbiology often (but not always) uses DNA sequence data.
- Data: Mutation can often be neglected in genetic profiling, but must be accounted for in forensic microbiology.
- Analysis: Mendelian analysis is the main tool used to interpret human genetic profiles, while statistical phylogenetics and population genetics are more important in forensic microbiology.

Phylogenetic analysis is concerned with establishing the evolutionary relationships that hold within a group of individuals. This can be done by exploiting the fact that related individuals tend to share similar characters which they have inherited from common ancestors. These characters can include both morphological and behavioral traits, but most phylogenetic analyses of

extant species (and even of some extinct species) now rely almost exclusively on genetic sequence data. The relationships between the individuals in a sample can be represented as a **phylogenetic tree** (also called a **genealogy**), which is a connected graph without cycles. The nodes in a phylogenetic tree correspond either to the sampled individuals (the **leaves** of the tree) or to common ancestors of groups of these individuals, while the branches in a tree correspond to lineages. A tree can either be **rooted**, in which case there is a single internal node that is the common ancestor of all of the individuals in the sample (the **most recent common ancestor**), or the tree can be **unrooted** if such a node is not designated. Rooted trees have a natural orientation, in which branches are regarded as directed edges that point away from the root and towards the leaves. A phylogenetic tree can encode two kinds of information: the topology of the tree shows groups of individuals that are more closely related to one another than to any other individual in the sample, while the branch lengths may indicate the absolute or relative time that has elapsed since a group of individuals last shared a common ancestor. A tree is said to be **binary** if every node has either one or three neighbors (two nodes are said to be neighbors if they are connected by a single branch): leaves have exactly one neighbor, while all other nodes have three. Phylogenetic trees are often, but not always, binary: for example, it is not always possible to resolve the order in which a rapid series of branching events has occurred, in which case these may be collapsed into a **polytomy**. One of the main challenges faced in phylogenetics is that even if we restrict attention to binary trees, the number of such trees grows more than exponentially rapidly as a function of the number of individuals in the sample. In particular, the number of unrooted binary trees with n labeled leaves is

$$(2n-5)!! = \frac{(2n-4)!}{(n-2)!2^{n-2}} = 1 \times 3 \times 5 \times \cdots \times (2n-5)$$

while the number of rooted binary trees is $(2n-3)!!$, which is even larger. For example, when $n = 20$, these numbers are approximately 2.22×10^{20} and 8.2×10^{21}, respectively, i.e., they are too large to allow exhaustive enumeration of all possible binary trees in the course of a phylogenetic analysis.

In part because of the combinatorial difficulty of the problem, many methods have been proposed for phylogenetic analyses. These include:

- **distance methods (e.g., neighbor-joining)**: these methods only consider the matrix of pairwise distances between taxa;
- **parsimony methods**: these methods try to find the tree that requires the least number of weighted changes in character states;
- **likelihood inference**: these methods try to find the tree that has the highest likelihood under a model of character evolution;
- **Bayesian phylogenetics**: these methods try to generate a sample of trees from the posterior distribution.

Because parsimony and likelihood methods require the optimization of an objective function over the space of all possible trees, it is usually not possible to implement these methods exactly when working with more than about 6 taxa. Instead, heuristic methods are used to examine a subsample of all possible trees that ideally will include, if not the optimal tree itself, then at least some tree that is 'close' in some sense to the optimal tree. A similar problem arises in Bayesian phylogenetics, where the posterior sample of trees may not be large enough to accurately approximate the true posterior distribution.

8.2 Substitution Processes

To do statistical phylogenetics, we need to be able to calculate the conditional probability of the observed data (in terms of character traits or genetic sequences) as a function of the evolutionary relationships between the sampled individuals. Usually this is done by adopting a stochastic model that describes how the characters in question evolve as they are transmitted along each branch of the tree. It is also usually assumed that evolutionary changes occur independently along different branches of the tree. While this assumption may not apply to characters subject to common selection pressures operating on different lineages, it is probably a reasonable approximation when modeling neutral evolution.

In molecular phylogenetics, the models used to describe the evolution of DNA and RNA sequences as they are transmitted from one generation to the next are called **substitution processes**. Although substitution and mutation are closely related, there is an important difference, which is that a mutation in a DNA sequence only leads to a substitution if it is transmitted from one generation to the next. Furthermore, although the two processes coincide at neutrally evolving sites, the rate of substitution at a non-neutral site may be less than or greater than the mutation rate at that site depending on whether natural selection impedes or favors the transmission of new mutations to the next generation. For example, there is a large body of evidence that the substitution rate in coding DNA is strictly less than the mutation rate and that this is because some mutations are deleterious and so are removed from populations by natural selection.

To describe the substitution processes that are commonly used in molecular phylogenetics, I first need to introduce some background material on stochastic processes and continuous-time Markov chains. Suppose that $E = \{e_1, e_2, \cdots, e_n\}$ is a finite set. An E-valued **continuous-time stochastic process** is simply a sequence of random variables $\{X_t, t \geq 0\}$ with the property that each variable X_t is a random variable with a value in E. The process is said to be continuous in time because the random variables X_t are indexed by a parameter that can be any non-negative real value. Here we interpret X_t as the state of a system of interest at time t.

A continuous-time stochastic process $(X_t; t \geq 0)$ is said to be a **continuous-time Markov chain** (CTMC) if it has the property that for any set of times $s < t < u$, the random variables X_s and X_u are conditionally independent given the value of X_t. This is called the **Markov property** and it means that if we know the current value of the process, say $X_t = x$, then we can predict the future values of the process without having to know how the process came to arrive at state x at time t. Since there are potentially infinitely many ways in which the process could have arrived at this particular state, Markov processes are much easier to work with than are stochastic processes that lack the Markov property. In particular, one can use the Markov property to show that any CTMC $X = (X_t; t \geq 0)$ with values in the set E can be described by a matrix Q, called the **rate matrix** of X, and by the **initial distribution** μ of the variable X_0. If E contains n different states, then the rate matrix $Q = (q_{ij})$ is an $n \times n$-dimensional matrix with entries $q_{ij} \geq 0$ that specify the transition rate from state e_i to state e_j when $j \neq i$. Also, by convention, the diagonal elements q_{ii} of the rate matrix are chosen so that the sum of the elements in each row is equal to 0: for each $i = 1, \cdots, n$, we set

$$q_{ii} = -\sum_{j \neq i} q_{ij}.$$

Notice that the absolute value of the diagonal element $|q_{ii}|$ is just the total rate at which the process transitions from state i to some other state.

Given the rate matrix Q and the initial distribution μ, the behavior of the chain X has the following simple description. We first choose a point from E at random according to the initial distribution μ and we set X_0 equal to this value. If this state is e_i, say, we then simulate an

exponentially-distributed random variable τ_0 with mean $1/|q_{ii}|$, which determines how long the process remains within its initial state. τ_0 is said to be the **holding time** in the initial state. Recall that the density of an exponential random variable with this mean is

$$p_{\tau_0}(t) = |q_{ii}|e^{-|q_{ii}|t}, \quad t \geq 0$$

and notice that the duration of this holding time will usually be inversely proportional to the total transition rate out of state e_i. At time τ_0, the chain jumps to a new state in E that is chosen according to the following probability distribution: given that the initial state is e_i, the new state is chosen to be e_j with probability

$$q_{ij}/|q_{ii}|.$$

In other words, the larger the transition rate from state e_i to e_j is, the more likely it is that the chain will move to e_j rather than to some other state. In particular, if $q_{ij} = 0$ for some j, then the probability that the chain moves to state e_j is 0. Furthermore, if we know the initial state of the chain, then the holding time in that state and the new state to which the chain jumps at the end of the holding time are independent of one another. This is a consequence of the Markov property.

The future behavior of the chain is determined similarly. Following the k'th jump, the chain will occupy a new state, say e_l, and the holding time in that state will be an exponentially-distributed random variable τ_k with mean $1/|q_{ll}|$. At the end of the holding time, the chain will jump to a new state e_m in E that is chosen according to the probability distribution with weights

$$q_{lm}/|q_{ll}|.$$

This state depends only on the state occupied by the chain immediately before the jump, but it is independent of both the holding times $\tau_0, \cdots, \tau_k$ and of the states occupied by the chain prior to the k'th jump. These assertions too are consequences of the Markov property. The construction described in this paragraph can be used to simulate Markov chains on a computer and is known as the Gillespie algorithm.

One of the most fundamental questions that we can ask about a Markov chain concerns the probability that the chain is in state e_j at time t given that it was in state e_i at time 0. This quantity is called the **transition probability** from state e_i to e_j at time t and the $n \times n$ matrix

$$P(t) = (p_{ij}(t))$$

is called the **transition matrix**. The transition probabilities can be found by solving the following system of first-order linear differential equations,

$$\dot{P}(t) = P(t)Q,$$

which are known variously as the **master equations** or the **forward Kolmogorov equations**. For fixed i, j, the corresponding equation has the form

$$\dot{p}_{ij}(t) = \sum_{k \neq j} q_{kj} \cdot p_{ik}(t) - |q_{jj}| \cdot p_{ij}(t).$$

An explicit solution can be found by exponentiating the rate matrix and is equal to

$$P(t) = e^{Qt} \equiv \sum_{n=0}^{\infty} \frac{1}{n!} Q^n t^n,$$

where $Q^n = Q \cdots Q$ is the usual matrix product.

Example 8.1. *Suppose that $X = (X_t; t \geq 0)$ is a CTMC with state space is $E = \{1, 2\}$ and that the rate matrix is*

$$Q = \begin{pmatrix} -a & a \\ b & -b. \end{pmatrix},$$

i.e., a is the transition rate from 1 to 2, while b is the transition rate from 2 to 1. Since there are only two states, we know that the transition probabilities out of each state must sum to 1:

$$\begin{aligned} p_{11}(t) + p_{12}(t) &= 1 \\ p_{21}(t) + p_{22}(t) &= 1. \end{aligned}$$

These can be used to simplify the Kolmogorov forward equations

$$\begin{aligned} \dot{p}_{11}(t) &= -ap_{11}(t) + bp_{12}(t) \\ &= b - (a+b)p_{11}(t) \\ \dot{p}_{22}(t) &= -bp_{22}(t) + ap_{21}(t) \\ &= a - (a+b)p_{22}(t), \end{aligned}$$

which can then be solved to give

$$P(t) = e^{Qt} = \frac{1}{a+b} \begin{pmatrix} b + ae^{-(a+b)t} & a - ae^{-(a+b)t} \\ b - be^{-(a+b)t} & a + be^{-(a+b)t}. \end{pmatrix}$$

In particular, if $a + b > 0$, then as t tends to infinity, this matrix tends to the limit

$$P(\infty) = \frac{1}{a+b} \begin{pmatrix} b & a \\ b & a \end{pmatrix},$$

which shows that after a large amount of time has elapsed, the chain 'forgets' its initial state and is found in state 1 with probability $b/(a+b)$ and state 2 with probability $a/(a+b)$.

The limiting probability distribution found in the preceding example is said to be a **stationary distribution** or **equilibrium distribution** for the Markov chain. In general, a probability distribution π on the state space E is said to be a stationary distribution for a Markov chain X with values in E if whenever the initial distribution is π, then the distribution of the chain at every time $t \geq 0$ is π. Every Markov chain with values in a finite state space E has at least one stationary distribution and some chains have more than one stationary distribution. (Take $a = b = 0$ in the previous example for an example of a chain with infinitely many stationary distributions.) Some Markov chains have a much stronger property, called **ergodicity**, which means that their distributions at large times t converge to a unique stationary distribution π as $t \to \infty$ no matter what the initial distribution is. This is true of the chain in the previous example whenever $a, b > 0$.

Substitution processes in molecular evolution are usually modeled using continuous-time Markov chains either on the set of four nucleotides $E = \{T, C, A, G\}$ or $E = \{U, C, A, G\}$ for DNA or RNA sequences, respectively, or on the set of 64 possible triplet codons. Codon-based models are important when modeling the evolution of coding sequences, but here we will concentrate on nucleotide-based models for DNA sequence evolution. We will also adopt an arbitrary order for the four bases, setting $e_1 = T, e_2 = C, e_3 = A, e_4 = G$ and noting that different authors use different conventions. Implicit in writing down such a model is the common assumption that substitution processes at different sites are independent of one another.

Example 8.2. *The Jukes-Cantor model is the simplest CTMC process used to model the neutral substitution process at a single site in a DNA molecule. It assumes that substitution rate from any one base to any other is a constant μ, so that the rate matrix is just*

$$Q = \begin{pmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{pmatrix},$$

while the transition matrices are given by

$$P(t) = e^{Qt} = \frac{1}{4}\begin{pmatrix} 1+3e^{-4\mu t} & 1-e^{-4\mu t} & 1-e^{-4\mu t} & 1-e^{-4\mu t} \\ 1-e^{-4\mu t} & 1+3e^{-4\mu t} & 1-e^{-4\mu t} & 1-e^{-4\mu t} \\ 1-e^{-4\mu t} & 1-e^{-4\mu t} & 1+3e^{-4\mu t} & 1-e^{-4\mu t} \\ 1-e^{-4\mu t} & 1-e^{-4\mu t} & 1-e^{-4\mu t} & 1+3e^{-4\mu t} \end{pmatrix}.$$

In particular, if we let $t \to \infty$, then we see that the Jukes-Cantor model is an ergodic Markov chain with stationary distribution $(1/4, 1/4, 1/4, 1/4)$, i.e., this model implies that all four bases occur at equal frequencies at equilibrium.

If we compare homologous sites in the genomes of two individuals that shared a common ancestor t units of time ago, then under the Jukes-Cantor model, the probability that the two sites are identical is $(1+3e^{-8\mu t})/4$, while the probability that they differ is $3(1-e^{-8\mu t})/4$. Now, suppose that we compare L homologous sites between these two individuals and assume that each site evolves at the same rate according to the Jukes-Cantor model and that the substitution processes at the different sites are independent. If we know the mutation rate μ (expressed in mutations per site per unit time), then this model can be used to estimate the divergence time between the two individuals. In particular, if there are d sites at which the two genotypes differ, then the likelihood function for t (treated as a parameter) is

$$L(t; d) = \left(\frac{3\left(1-e^{-8\mu t}\right)}{4}\right)^d \left(\frac{1+3e^{-8\mu t}}{4}\right)^{L-d},$$

while the log-likelihood function is

$$l(t; d) = C + d\log\left(1-e^{-8\mu t}\right) + (L-d)\log\left(1+3e^{-8\mu t}\right),$$

where C is a constant that does not depend on t. The maximum likelihood estimate of t can be found by differentiating the log-likelihood function with respect to t and setting this equal to 0. This gives the equation

$$0 = d\frac{8e^{-8\mu t}}{1-e^{-8\mu t}} + (L-d)\frac{-24e^{-8\mu t}}{1+3e^{-8\mu t}},$$

which can be solved explicitly. (To do so, first let $x = e^{-8\mu t}$ and solve for x, and then solve for t.) After some algebra, we find that the maximum likelihood estimate of the divergence time is equal to

$$\hat{t} = \frac{-1}{8\mu}\log\left(1-\frac{4}{3}\frac{d}{L}\right),$$

provided that $d/L < 3/4$. When $d/L \ll 1$, then the estimated divergence time is approximately linear in the proportion of sites that differ,

$$\hat{t} \approx \frac{1}{6\mu}\frac{d}{L},$$

but this approximation breaks down as d/L increases due to the occurrence of multiple mutations at sites. In particular, some fraction of substitutions will go unobserved due to reverse mutations

that restore the ancestral nucleotide, e.g., T may mutate to A which can then mutate back to T. This phenomenon is known as saturation and makes it difficult to accurately estimate large divergence times (although lower bounds can be given).

Although the Jukes-Cantor model is important for historical reasons, the assumption that all four bases mutate at equal rates to all other bases is known to be too simplistic a model for the evolution of most DNA sequences. For this reason, more general models have been introduced and the one that has perhaps seen the widest application in molecular phylogenetics is called the general time reversible (GTR) model, which has the following rate matrix

$$Q = \begin{pmatrix} \circ & \pi_C\alpha & \pi_A\beta & \pi_G\gamma \\ \pi_T\alpha & \circ & \pi_A\delta & \pi_G\epsilon \\ \pi_T\beta & \pi_C\delta & \circ & \pi_G\eta \\ \pi_T\gamma & \pi_C\epsilon & \pi_A\eta & \circ \end{pmatrix},$$

where $(\pi_T, \pi_C, \pi_A, \pi_G)$ is the stationary distribution for this model, $\alpha, \beta, \gamma, \delta, \epsilon, \eta$ are six free parameters (to be estimated from data), and the circles on the diagonal indicate that these entries should be chosen so that the row sums are zero. Although an analytical expression for the transition matrix $P(t) = e^{Qt}$ can be written down for the GTR process, it is very complicated and these probabilities are often calculated by numerical evaluation of the matrix exponential. The process is said to be **time reversible** because it satisfies the following condition,

$$\pi(e_i) \cdot \mathbb{P}(X_{t+s} = e_j | X_t = e_i) = \cdot\pi(e_j)\mathbb{P}(X_{t+s} = e_i | X_t = e_j),$$

which says that, at equilibrium, the transition probabilities of the Markov chain do not change if we reverse the direction of time. While this condition has no biological justification, it is a convenient assumption because it allows us to calculate the likelihoods of unrooted trees by assigning an arbitrary direction to the flow of time. In fact, the GTR model is the most general class of CTMC on the state space $E = \{T, C, A, G\}$ with this property and several other substitution processes commonly encountered in molecular phylogenetics, including the Jukes-Cantor model, are themselves special cases of the GTR. One model of particular importance is called the HKY85 model (after the three authors who proposed it in 1985, Hasegawa, Kishino and Yano). It accounts for the different rates at which transitions and transversions occur and has a rate matrix that can be expressed as

$$Q = \begin{pmatrix} \circ & \pi_C\kappa & \pi_A & \pi_G \\ \pi_T\kappa & \circ & \pi_A & \pi_G \\ \pi_T & \pi_C & \circ & \pi_G\kappa \\ \pi_T & \pi_C & \pi_A\kappa & \circ \end{pmatrix},$$

where the parameter κ is interpreted as the transition/transversion rate ratio. In humans, typical values of κ are around 2 with some variation between coding and non-coding loci.

8.3 Branching Processes and Coalescents

In Bayesian phylogenetics, inference is based on the posterior distribution of phylogenetic trees given by the formula:

$$p(\mathcal{T}, \Theta | D) = p(D | \mathcal{T}, \Theta) \frac{p(\mathcal{T}, \Theta)}{p(D)} \tag{8.1}$$

where D is the sequence data, $\mathcal{T}$ is a phylogenetic tree, and Θ is a set of model parameters such as the substitution rate matrix. The conditional probability $p(\mathcal{T}, \Theta | D)$ is said to be the

posterior distribution of $\mathcal{T}$ and Θ and it depends on both the **likelihood** of the data given by $p(X|\mathcal{T},\Theta)$ and also on the **prior distribution** $p(\mathcal{T},\Theta)$ of the tree and model parameters. The unconditional probability $p(D)$ can be regarded as a normalizing constant and usually does not play a role in inference.

It is often the case that the set of model parameters can be decomposed into two disjoint sets,

$$\Theta = \Theta_T \cup \Theta_S,$$

where Θ_T contains parameters that govern the prior distribution of the phylogenetic tree and Θ_S contains any parameters that are independent of the tree. We can then write the joint prior distribution on $\mathcal{T}$ and Θ as the product

$$p(\mathcal{T}|\Theta_T)p(\Theta_T,\Theta_S), \tag{8.2}$$

while the likelihood function can often be expressed as

$$p(D|\mathcal{T},\Theta_S), \tag{8.3}$$

i.e., once we condition on the tree, the data only depends on the parameters in Θ_S. If the phylogenetic tree is the only object of interest, then one can obtain the posterior distribution of $\mathcal{T}$ alone by integrating over Θ in equation (8.1):

$$p(\mathcal{T}|D) = \frac{1}{p(D)} \int p(D|\mathcal{T},\Theta_S)p(\mathcal{T}|\Theta_T)p(\Theta_T,\Theta_S)d\Theta_T d\Theta_S. \tag{8.4}$$

It is almost always the case that this integral cannot be evaluated analytically or even numerically and then Monte Carlo methods are used to estimate its value.

Our purpose in this section is to describe some of the prior distributions on phylogenetic trees that have found to be useful in Bayesian phylogenetics. Recall that the role of the prior distribution in Bayesian statistics is to summarize any existing knowledge or beliefs that we may have about the unknown elements of a statistical model before we examine the new data that is to be explained by the model. In Bayesian phylogenetics, this prior information could take into account alternative data sources such as fossil evidence, but often it comes from mathematical theories that describe the distribution of trees that are expected under a particular model of evolution or demography. Here we consider two classes of models known as branching processes and coalescent processes.

8.3.1 Branching Processes

A continuous-time branching process is a continuous-time Markov chain $X = (X_t; t \geq 0)$ that describes the growth of a population in which each individual reproduces independently of the rest of the population. The variable X_t specifies the number of individuals alive at time t and the population is said to be extinct at time t if $X_t = 0$. The behavior of a branching process depends on two parameters: the branching rate $\lambda > 0$ and the offspring distribution ν which is a probability distribution on the non-negative integers. Gillespie's algorithm can be used to simulate a branching process in the following manner. If $X_t = n$, choose one of the n individuals at random and generate both an exponentially distributed random variable τ with mean $1/(\lambda n)$ and an integer-valued random variable ξ with distribution ν. τ and ξ are assumed to be independent of each other and of the past history of the population. Then, at time $t+\tau$, replace the chosen individual by its ξ descendants and set $X_{t+\tau} = n - 1 + \xi$. Although X itself is an integer-valued process that only records the population size at different times, notice that our description of this process also tells us how to construct the random tree that describes the genealogical relationships between the individuals in the population.

The **Yule process** (also called the Yule-Ferry or the pure birth process) is an important special case of the general continuous-time branching process in which every reproduction event results in the birth of exactly two new individuals. In this case the tree is binary and the population increases by one at each branching event. In fact, one can use the forward equations to solve for the probability distribution of the number of individuals alive at time t. For example, if the population is initially founded by a single individual, then $X_0 = 1$ and it can be shown that X_t is a geometrically-distributed random variable with mean $e^{\lambda t}$, i.e.

$$\mathbb{P}(X_t = n) = e^{-\lambda t}\left(1 - e^{-\lambda t}\right)^{n-1}, \quad n \geq 1.$$

In Bayesian phylogenetics, the Yule process as well as other more complicated branching processes are sometimes used to specify the prior distribution on the phylogenetic tree that describes the evolutionary relationships amongst a group of species. In other words, the prior probability of a particular tree $\mathcal{T}$ is set equal to the probability density of that tree under the chosen model. When the Yule process is used, there is one tree-related parameter λ which determines the branching rate. The value of λ can either be estimated from the data or be given a prior distribution itself as in equation (8.2).

8.3.2 Coalescent Processes

The phylogenetic tree of a sample of individuals from a single randomly-mating population of constant size has a well-studied description in terms of a class of continuous-time Markov chains known as **coalescent processes**. Recall that in the Wright-Fisher model (cf. Section 3.2), the parents of the N individuals alive in generation $t + 1$ are determined by sampling N times at random and with replacement from amongst the individuals alive in generation t. Although we were previously interested in this model for what it could tell us about the effect of random genetic drift on the amount of genetic variation maintained in a finite population, the construction described in the previous sentence can also be used to determine the genealogy of a random sample of individuals taken from a given generation. For example, if we sample just two individuals from generation t, then the probability that they share a common ancestor in generation $t-1$ is $1/N$. If they don't share a common ancestor in the preceding generation, then the probability that they share a common ancestor two generations ago is $(1-1/N)\cdot 1/N$, while the probability that they first share a common ancestor k generations ago is

$$\left(1 - \frac{1}{N}\right)^{k-1} \cdot \frac{1}{N}.$$

In other words, if we write τ_2 for the number of generations in the past when the two individuals first shared a common ancestor, then τ_2 is a geometrically-distributed random variable with mean N. The two lineages of ancestors to these individuals are said to **coalesce** (i.e., come together) at time τ_2 and τ_2 is called the coalescent time for the two lineages.

If three individuals are sampled at random from the same population, then the probability that all three share a common ancestor in the preceding generation is N^{-2}. When N is large, this probability is much smaller than the probability that exactly two of the individuals share a common ancestor in that generation, which can be written as

$$\frac{3}{N} + O\left(\frac{1}{N^2}\right).$$

The factor of three in the first term accounts for the fact that there are three distinct pairs of individuals that could share a common ancestor in the preceding generation. Similar statements apply to earlier generations and it can be shown that when N is large, the genealogy of a random

sample of three individuals is much more likely to be a binary tree than to contain a complex coalescent event involving the simultaneous merger of more than two ancestral lineages. In fact, this remains true even when we consider samples containing a larger number of individuals, say n, provided that n is much smaller than N.

A simple approximation for the genealogy of a random sample of n individuals can be derived by measuring time in units of N generations and taking the limit as N tends to infinity with n fixed. It can be shown that only pairwise coalescent events occur and that the time to the next coalescent event when there are k lineages in the tree is exponentially distributed with mean equal to

$$\binom{k}{2}^{-1} = \frac{2}{k(k-1)}.$$

In fact, the genealogy of a random sample of n individuals can be generated in the following manner. We first generate an exponentially-distributed random variable τ_n with mean $\binom{n}{2}^{-1}$. At that time, two of the lineages in the tree are chosen uniformly at random and merged into a single lineage. We next generate an independent exponentially-distributed random variable τ_{n-1} with mean $\binom{n-1}{2}^{-1}$ and then choose two of the remaining $n-1$ lineages in the tree at random and merge them into a single lineage. This process continues until only a single lineage is left and the time, say T_{MRCA}, when this first occurs is said to be the **time to the most recent common ancestor** (TMRCA) of the sample. A formal description of this process was first given by J.F.C. Kingman in 1983 and consequently the process has come to be known as **Kingman's coalescent**. Notice that the expected value of the TMRCA is equal to

$$\begin{aligned}
\mathbb{E}\left[T_{MRCA}\right] &= \mathbb{E}\left[\tau_n + \tau_{n-1} + \cdots + \tau_2\right] \\
&= \sum_{k=2}^{n} \frac{2}{k(k-1)} \\
&= 2\sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k}\right) \\
&= 2\left(1 - \frac{1}{n}\right),
\end{aligned}$$

where we recall that time is measured in units of N generations. Because the rate of coalescence is a quadratic function of the number of lineages in the tree, branches near the top of a coalescent tree tend to be much shorter than branches close to the root.

Although we obtained Kingman's coalescent by way of the Wright-Fisher model, Kingman and others have shown that this particular process provides a surprisingly accurate approximation for the distribution of genealogies obtained from a much larger class of models of genetic drift provided that the following conditions are satisfied:

- the population size N is large and constant;
- the sample size n is much smaller than N;
- the population is randomly mating;
- the variance in reproductive success is not too large;
- the individuals are selectively equivalent (neutral evolution).

Various generalizations of the coalescent have been derived which allow some of these conditions to be relaxed. Provided that the population never becomes too small (i.e., no bottlenecks), variable population sizes can be accommodated by a simple change in the rate of coalescence.

In fact, by stipulating a parametric model of population size change such as exponential growth, e.g.,

$$N_t = N_0 e^{\gamma t},$$

one can use Bayesian phylogenetic methods to infer the demographic history of a population from a sample of DNA or RNA sequences. Spatial structure is arguably the most important feature of real populations that has not yet been adequately addressed by coalescent-based phylogenetic analyses, but some progress in this direction has been made in recent years (cf. Lemey et al. 2009, PLoS Comput. Biol.).