Mutation-Drift Balance

February 15, 2015
1 Infinite Alleles Model

1.1 Motivation

Our initial study of the Wright-Fisher model demonstrated that, in the absence of mutation and immigration, genetic drift will ultimately result in the loss of all genetic variation from a population. Furthermore, we showed that the rate at which variation is lost is inversely proportional to the population size. However, real populations often harbor genetic variation, which ultimately must arise through mutation. Thus, in these notes, we will investigate two models that incorporate both mutation and genetic drift. Our objective is to understand how variation is affected when both processes act within a single population.

Our first model is known as the infinite alleles model and was introduced by M. Kimura and J. Crow in a paper published in 1964. This model makes the following three assumptions:

1. Reproduction occurs according to the Wright-Fisher model, i.e., the population size is constant, there are non-overlapping generations, and each chromosome chooses its ancestor uniformly at random and with replacement from the previous generation. Throughout these notes we will assume that the population is diploid and contains $2N$ chromosomes.

2. Each chromosome transmitted to the next generation mutates with probability $\mu$.

3. Each mutation results in the creation of a new allele which has not previously occurred within the population.

It is the last assumption that accounts for the name of the model. There are several convenient schemes for labelling the alleles that appear in the model. One approach is to assign integer labels that reflect the order of appearance, e.g., the wild type allele is labelled 0, the allele generated by the first mutation is labelled 1, etc. Alternatively, we can assign each allele a unique label by sampling a real number in $[0,1]$ uniformly at random; this second approach is particularly convenient when studying the infinite population size limit of the infinite alleles model which is a type of measure-valued diffusion known as the Fleming-Viot process (Fleming & Viot, 1979).

The infinite alleles assumption was motivated by three considerations. First, although we might like to model the evolution of an actual DNA sequence, the geometric growth in the number of alleles as a function of sequence length makes this intractable. Indeed, to exactly model variation in a sequence of length $L$, we would have to keep track of the frequencies of $4^L - 1$ alleles. On the other hand, because nucleotide mutation rates are typically very small - on the order of $10^{-8} - 10^{-10}$ mutations per site per generation - the probability of multiple independent origins for any allele segregating in a population of size less than the inverse of the mutation rate is also very small, especially when $L$ is large. The third motivation relates to the manner in which genetic variation was measured in the 1960’s and 1970’s before DNA sequencing became tractable. At that time, variation within populations was often studied using gel electrophoresis of proteins which can detect different forms of a protein (known as allozymes) that differ in length or charge. In particular, although these methods were able to reveal polymorphism at the protein level, they did not reveal either the underlying genetic changes responsible for this variation or the relationships between the different allozymes.

1.2 Sample distributions: simulations

Suppose that we sample $n$ chromosomes at random from a population governed by the infinite alleles model and we wish to use this data to characterize the amount of variation contained in
the population itself. There are several summary statistics that we can calculate from our data. One of these is the number of different alleles sampled, which we will denote $K_n$. Alternatively, we could calculate the allelic partition of our data, which specifies the number of alleles that occur just once in the sample (singletons), the number of alleles that occur exactly twice in the sample, etc. To this end, let $a_i$ be the number of alleles that occur exactly $i$ times in the sample and observe that the following two identities will always be satisfied

\[ \sum_{i=1}^{n} a_i = K_n \]
\[ \sum_{i=1}^{n} a_i \cdot i = n. \]

Our objective in this section is to characterize the sampling distributions of these two statistics from a population that is at equilibrium.

Before we state any analytical results, let us consider how we could use simulations to investigate this question. Here I will describe five different approaches.

**Approach 1: Forward simulations.** One possibility would be to simulate the infinite alleles model forward in time for a large number of generations and then sample $n$ chromosomes from the population at the end of each simulation. Since genetic drift operates on a time scale proportional to the size of the population, we would have to run each simulation for order $O(2N)$ generations, say $10N$ to be concrete. Furthermore, since we have to generate $4N$ random variables in each generation ($2N$ to choose the parents and another $2N$ to assign mutations), this approach will require that roughly $40N^2$ random variables are generated per simulation. We will then have to repeat this procedure a large number of times, say at least 100, to obtain estimates of the sampling distributions of our summary statistics.

**Approach 2: Coalescent with mutation (I).** A much more efficient approach is to simulate the genealogy of a sample of size $n$ and then assign mutations to the resulting tree. Indeed, because our model is neutral, i.e., fecundity and survival are independent of an individual’s genotype, the marginal distribution of sample genealogies is given exactly by Kingman’s coalescent. Furthermore, because mutations occur at a constant rate, the number of mutations that occur along a branch of the genealogy of length $\tau$ (in generations) is Poisson-distributed with mean $\mu \cdot \tau$. In fact, the process of mutations along each branch of the genealogy is just a Poisson process with intensity of $\mu$. Notice that the computational burden of this approach is essentially independent of $N$.

**Approach 3: Coalescent with mutation (II).** An alternative approach is to assign mutations to the genealogy as the coalescent is being simulated backwards in time. This leads to a continuous-time Markov chain in which the following transitions can occur. First, as in the ordinary coalescent, each pair of lineages coalesces at rate $1/2N$. Secondly, mutations occur along each lineage at rate $\theta/2$ per generation. For reasons that will become apparent later, it is convenient to measure time in units of $2N$ generations and to introduce a new parameter $\theta = 4N\mu$ which is called the scaled mutation rate. With these conventions, pairs of lineages coalesce at rate 1, while mutations occur along each lineage at rate $\theta/2$.

**Approach 4: Coalescent with killing.** Because the infinite alleles model does not retain information about the genetic relationships between the different alleles that appear, we only need to simulate lineages back until we encounter the most recent mutation or until we reach the root of the tree. This observation leads to another continuous-time Markov chain that is sometimes described as a ‘coalescent with killing’. In this process, each pair of lineages still coalesces at rate 1, while, in addition, each lineage is killed at rate $\theta/2$. When a lineage is killed, it is assigned a novel allele and then removed from the coalescent. The process is continued...
until either all but one lineage has been killed, the remaining lineage being assigned a unique mutation, or until we reach the root of the tree, which is also assigned a unique mutation. The result of this simulation is a set of random trees, i.e., a random forest, where each tree consists of a set of lineages that share the same unique allele. In fact, it suffices to simulate just the jump chain of this process, which has the following transition probabilities. First, note that the total rate of transitions when there are \( n \) (live) branches is \( \binom{n}{2} + n\theta/2 \). It follows that the probability of a coalescent event in the jump chain is

\[
P_\theta(\text{coalescence}) = \frac{\binom{n}{2}}{\binom{n}{2} + n\theta/2} = \frac{n-1}{n-1+\theta},
\]

in which case one of the \( \binom{n}{2} \) pairs of lineages is chosen uniformly at random and coalesced, reducing the number of lineages to \( n-1 \). The other possibility is that one of the \( n \) lineages is killed, which occurs with probability

\[
P_\theta(\text{killing}) = \frac{\theta}{n-1+\theta},
\]

in which case the number of lineages is again reduced to \( n-1 \). Notice that we require exactly \( n-1 \) random variables to simulate the coalescent with killing for a sample of size \( n \).

**Approach 5: Hoppe’s urn.** Instead of simulating the coalescent with killing backwards in time, starting with the \( n \) sampled chromosomes and running the process until there is only one remaining lineage, we can also simulate this process from the root up to the leaves. The resulting process is known as Hoppe’s urn and is a discrete-time Markov chain with the following transition probabilities. Suppose that there are \( k \) branches in the random forest. Looking forwards in time, there are two possibilities. Either one of the \( k \) branches splits, which happens with probability

\[
P_\theta(\text{split}) = \frac{k}{k+\theta},
\]

in which case a randomly chosen branch is assigned two descendants, both of which inherit their ancestor’s allele, or a new branch is created and assigned a unique allele, which happens with probability

\[
P_\theta(\text{migration}) = \frac{\theta}{k+\theta}.
\]

Notice that no matter what type of event occurs, the number of branches always increases by one. Since we initiate the process at time 0 with zero branches, the very first event is certain to be an immigration. The process is halted once there are \( n \) branches in the random forest, each of which will have been assigned an allele that it shares with any other branches that are connected to it.

### 1.3 Sampling distributions: theory

Hoppe’s urn is important not only because it can be used to carry out efficient simulations, but also because it can be used to derive explicit analytical results about the sampling distribution of the infinite alleles model. We begin with a result describing the asymptotic properties of the number of alleles in a sample, \( K_n \), when \( n \) is large. Recall that two sequences of real numbers, \( a_n \) and \( b_n \), are said to be asymptotically equivalent as \( n \to \infty \), written \( a_n \sim b_n \), if

\[
\lim_{n \to \infty} \frac{a_n}{b_n} = 1.
\]

(Notice that it is not necessarily true that \( a_n - b_n \to 0 \) even if \( a_n \sim b_n \)).
Theorem 1. For fixed $\theta$, the mean and the variance of $K_n$ are asymptotically equivalent to $\theta \ln(n)$ as $n \to \infty$, i.e.,

$$E[K_n] \sim \theta \ln(n) \quad \text{and} \quad \text{Var}(K_n) \sim \theta \ln(n).$$

In addition, $K_n$ is asymptotically normally distributed in the sense that for every real number $x$,

$$\lim_{n \to \infty} P_{\theta} \left( \frac{K_n - E[K_n]}{\sqrt{\text{Var}(K_n)}} \leq x \right) = P(Z \leq x)$$

where $Z$ is a standard normal random variable.

Proof. Imagine running Hoppe’s urn until we have accumulated $n$ lineages and let $\eta_i$ be a Bernoulli random variable which is equal to 1 if the $i$'th event is a mutation and 0 otherwise. Since there are $i - 1$ lineages just before the $i$'th event, it follows that

$$P_{\theta}(\eta_i = 1) = \frac{\theta}{\theta + i - 1}.$$ 

Furthermore, all of the variables $\eta_1, \cdots, \eta_n$ are independent and $K_n = \eta_1 + \cdots + \eta_n$ since each mutation introduces a new allele into the sample. This shows that

$$E[K_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1},$$

which in turn implies that $E[K_n] \sim \theta \ln(n)$ since the right-hand side can be interpreted as a left Riemann sum for the integral

$$\theta \int_{\theta}^{n+\theta} \frac{dx}{x} = \theta (\ln(n + \theta) - \ln(\theta)) \sim \theta \ln(n).$$

Similarly, since the variance of a sum of independent random variables is equal to the sum of the variances, it follows that

$$\text{Var}(K_n) = \sum_{i=1}^{n} \frac{\theta(i - 1)}{(\theta + i - 1)^2} \sim \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \sim \theta \ln(n),$$

where we have used the fact that $(i - 1)/(\theta + i - 1) \sim 1$ as $i \to \infty$.

Finally, since the variables $\eta_1, \eta_2, \cdots$ are independent and

$$\lim_{n \to \infty} \frac{1}{\text{Var}(K_n)} \max_{1 \leq i \leq n} \text{Var}(\eta_i) = 0,$$

the Feller-Lindeberg central limit theorem implies that $K_n$, suitably normalized, is asymptotically normal as $n \to \infty$.

This theorem has two important consequences. First, it says that the mean number of alleles in a sample of size $n$ is linearly dependent on $\theta$, but only logarithmically dependent on the sample size. In other words, on average, the number of alleles in a sample grows very slowly with the sample size. Secondly, although we can use the statistic $K_n / \ln(n)$ to estimate $\theta$, the variance of this estimator is approximately

$$\text{Var}(K_n / \ln(n)) \sim \frac{\theta}{\ln(n)},$$
which decreases to 0 very slowly as \( n \to \infty \). The reason for this behavior is that each additional sample is not independent of the chromosomes already sampled. Instead, as \( n \) increases, every additional chromosome sampled is likely to be closely related to one of the chromosomes already in our sample and therefore contributes a decreasing amount of information concerning the value of \( \theta \).

Because \( K_n \) is a very crude summary of the full data, we might hope that more efficient estimators of \( \theta \) could be formulated that use all of the information contained in the data. Later we will see that this is not possible. However, our first step is to characterize the exact distribution of the allelic partition of a random sample. This distribution is known as Ewens’ sampling formula, which has also found application to certain problems arising in Bayesian nonparametric statistics and machine learning.

**Theorem 2.** Suppose that \( A_n \) is the allelic partition arising when \( n \) chromosomes are sampled at random from a population governed by the infinite alleles model. Then the distribution of \( A_n \) is given by the expression

\[
\Pr(\theta(A_n = (a_1, \ldots, a_n)) = \frac{n!}{\theta(n)} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!},
\]

where \((a_1, \ldots, a_n)\) is a vector of non-negative integers with \( a_1 + 2a_2 + \cdots + na_n = n \) and the Pochammer symbol \( \theta(n) = \theta(\theta + 1) \cdots (\theta + n - 1) \) denotes the rising factorial.

**Proof.** Once again, we use Hoppe’s urn to generate random samples and we verify the ESF by induction on \( n \). When \( n = 1 \), there is exactly one allelic partition \( a_1 = 1 \) and the ESF correctly asserts that this partition has probability 1. Now suppose that the ESF holds for \( n - 1 \) and let \( a = (a_1, \ldots, a_n) \) be a partition of a sample of size \( n \). By the law of total probability, we know that

\[
\Pr(\theta(A_n = a) = \sum_{a'} \Pr(\theta(A_{n-1} = a') \cdot \Pr(\theta(A_n = a|A_{n-1} = a')),
\]

where the sum is over all the partitions \( a' = (a'_1, \ldots, a'_{n-1}) \) of a sample of size \( n - 1 \) that can give rise to the partition \( n \) and \( \Pr(\theta(A_n = a|A_{n-1} = a')) \) is the transition probability of moving from \( a' \) to \( a \) under Hoppe’s urn. There are two cases. One possibility is that the \( n \)’th event involves a mutation, in which case \( a' = (a_1 - 1, a_2, \ldots, a_n) \) and

\[
\Pr(\theta(A_n = a|A_{n-1} = a')) = \frac{\theta}{\theta + n - 1}.
\]

The other possibility is that the \( n \)’th event involves a duplication of one of the lineages already present. If this involves an allele that was present in \( i \) copies at time \( n - 1 \), then \( a'_i = a_i + 1 \) and \( a'_{i+1} = a_i - 1 \), and the probability is

\[
\Pr(\theta(A_n = a|A_{n-1} = a')) = \frac{(a_i - 1) \cdot i}{\theta + n - 1},
\]

since there are \( a'_i \cdot i \) lineages at time \( n - 1 \) that carry alleles represented by \( i \) copies. By substituting these transition probabilities into the law of total probability and using ESF to calculate the probabilities \( \Pr(\theta(A_{n-1} = a')) \), we can show that \( \Pr(\theta(A_n = a) \) also satisfies the ESF. See the proof in Durrett for the details. This completes the induction.

By setting \( n = 2 \), Ewens’ sampling formula can be used to calculate the expected heterozygosity of a population governed by the infinite alleles model. Recall that the expected heterozygosity of a population is defined to be the probability that two chromosomes sampled at random carry
two distinct alleles. Since this will be true if and only if the allelic partition of the sample has $a_1 = 2$ and $a_2 = 0$, it follows that

$$
\bar{H} = \mathbb{P}_\theta(a_1 = 2, a_2 = 0) = \frac{\theta}{\theta + 1} = \frac{4N\mu}{1 + 4N\mu}.
$$

In fact, this result can also be derived directly from the coalescent with mutation by observing that the sample will contain two different alleles if and only if the first event in the genealogy is a mutation, which happens with probability $\theta/(\theta + 1)$. The resulting expression shows that the expected heterozygosity is an increasing function of both the mutation rate $\mu$ and the population size $N$. This is unsurprising since an increase in the mutation rate means that new alleles will be introduced into the population that much more rapidly, while an increase in the population size reduces the rate at which genetic variation is reduced by genetic drift. Moreover, $\bar{H}$ is a saturating function of $\theta$ which is equal to half of its maximum possible value ($\bar{H}_{\text{max}} = 1$) when $\theta = 1$.

Another important observation is that because the sampling distribution of the allelic partition depends only on the composite parameter $\theta$, the individual parameters $N$ and $\mu$ are not identifiable, i.e., there is no way to separately estimate the population size and the mutation rate under the infinite alleles model from a sample of chromosomes collected at a single time. Of course, if we have independent information that specifies either $N$ or $\mu$, then we can use an estimate of $\theta$ derived under this model to obtain an estimate of the other parameter. Alternatively, if we have dated samples which were collected at different times, then we can use this information to separately estimate both parameters. This is one of the reasons why ancient DNA is of such interest.

1.4 Sufficiency of $K_n$

Our next objective is to show that under the infinite alleles model $K_n$ contains all of the information relevant to the estimation of $\theta$. More formally, we will show that $K_n$ is a sufficient statistic for $\theta$, which is explained in the next definition.

**Definition 1.** Suppose that a random variable $X$ has distribution $\mathbb{P}_\theta$, where $\theta \in \Theta$ is a parameter. A statistic $T(X)$ is said to be a **sufficient statistic for** $\theta$ if for every $x$ and every $t$ the conditional probability

$$
\mathbb{P}_\theta(X = x | T(X) = t)
$$

does not depend on $\theta$.

In other words, by conditioning on a sufficient statistic, we remove any relationship between the full data and the unknown parameter. In particular, if we can show that the Ewens’ sampling formula depends on $\theta$ only through $K_n$, then it will follow that $K_n$ is a sufficient statistic for $\theta$. To this end, we will introduce yet another discrete-time Markov chain, $\Pi = (\Pi_n; n \geq 1)$, related to Hoppe’s urn and sometimes known as the **Chinese restaurant process**, which will take values in the set of permutations. Although $\Pi$ may appear to be even more complicated than Hoppe’s urn, it will lead to a simple expression for the exact sampling distribution of $K_n$. (In contrast, Theorem 1, stated above, only describes the asymptotic behavior of this distribution.) We begin by recalling some facts about permutations.

Recall that a permutation of the set $\{1, \cdots, n\}$ is a map $\pi$ which is both one-to-one and onto. For example, if we are given a deck of $n$ cards, labeled in order of appearance (say from the top of the deck down to the bottom), then any permutation $\pi$ corresponds to a way of reshuffling the deck, so that the card originally occupying the $i$’th position is moved to the $\pi(i)$’th position. In particular, we can represent any permutation $\pi$ by listing the values $(\pi(1), \cdots, \pi(n))$, e.g., if
\(n = 6\), then the permutation \(\pi = (5, 4, 1, 2, 3, 6)\) maps 1 to 5, 2 to 4, and so forth. Because the set \(\{1, \cdots, n\}\) is finite, repeated application of a permutation \(\pi\) will eventually map every point back to its starting position. For example, with \(\pi\) as above, repeated application of \(\pi\) will map 1 to 5, then to 3, and then back to 1, which can be represented in longhand by writing \(1 \to 5 \to 3 \to 1\) or, more succinctly, by writing (153), where each number is mapped to the number to its right and the last number in the list is mapped to the first. In this case, we say that (153) is a cycle of the permutation. In fact, this notation is not unique, since (153), (315) and (531) all represent the same cycle. Further inspection of \(\pi\) reveals two additional cycles, (24) = (42) and (6), i.e., 6 = \(\pi(6)\) is a fixed point of the permutation, and in fact we can represent \(\pi\) itself as a product of its cycles, i.e., \(\pi = (531)(42)(6)\). This representation also is not unique, since we can list the cycles in any order, e.g., \(\pi = (42)(6)(531)\) is equally valid. What is important is that each integer \(i \in \{1, \cdots, n\}\) belongs to one and exactly one cycle, which follows from the fact that \(\pi\) is one-to-one.

The Chinese restaurant process can be seen as an enriched version of Hoppe’s urn which uses partitions to represent the sequence of events that have occurred up to each time \(n\). Suppose that the balls in the urn are numbered in order of appearance. Since each event will result in the addition of new ball to the urn, we can identify the \(n\) balls in the urn at time \(n\) with the set \(\{1, \cdots, n\}\). To associate a permutation \(\Pi_n\) with this set, we will assume that each time the black ball is chosen, the new ball that is added to the urn starts a new cycle. On the other hand, if the color of the \(n\)’th ball was determined by choosing the \(k\)'th ball, then \(n\) will be inserted into the cycle containing \(k\) immediately to the left of \(k\). In this way, each cycle in \(\Pi_n\) will correspond to a group of balls that share the same color and we can use \(\Pi_n\) to reconstruct the sequence of events that happened up to time \(n\). Our next result shows that the distribution of \(\Pi_n\) depends only on the number of cycles in the permutation.

**Theorem 3.** If \(\pi\) is a permutation of \(\{1, \cdots, n\}\) with \(k\) cycles, then

\[
P_\theta(\Pi_n = \pi) = \frac{\theta^k}{\theta(n)}.
\]

**Proof.** Given \(\pi\), let \(\xi_i = \theta\) if the \(i\)'th event resulted in the creation of a new cycle \((i)\) and \(\xi_i = 1\) if \(i\) was added to an existing cycle. If \(\pi\) has \(k\) cycles, then exactly \(k\) of the \(\xi_i\) will be equal to \(\theta\), while the rest will be equal to 1. Consequently,

\[
P_\theta(\Pi_n = \pi) = \prod_{i=1}^{n} \frac{\xi_i}{\theta - 1 + \theta} = \frac{\theta^k}{\theta(n)},
\]

which completes the proof. \(\Box\)

Since the number of cycles in the random permutation \(\Pi_n\) is equal to the number of alleles (or colors) in the corresponding sample from Hoppe’s urn, the following theorem is an immediate consequence of Theorem 4.

**Theorem 4.** The distribution of \(K_n\) under the infinite alleles model is given by

\[
P_\theta(K_n = k) = |S^k_n| \cdot \frac{\theta^k}{\theta(n)},
\]

where \(|S^k_n|\) is the number of permutations of \(\{1, \cdots, n\}\) with exactly \(k\) cycles.

The numbers \(|S^k_n|\) play an important role in combinatorics and are known as Stirling numbers of the first kind. Although they cannot be expressed in closed form, they do satisfy a simple first-order recurrence relation:

\[
|S^k_n| = (n - 1)|S^k_{n-1}| + |S^{k-1}_{n-1}|.
\]
To see this, notice that each permutation \( \pi \) of \([1, \cdots, n]\) with \( k \) cycles can be constructed either by choosing a permutation \( \sigma \) of \([1, \cdots, n-1]\) with \( k-1 \) cycles and adding \((n)\) as an extra cycle (a fixed point) or by choosing a permutation \( \sigma \) of \([1, \cdots, n-1]\) with \( k \) cycles and inserting the element \( n \) between one of the elements \( 1 \leq j \leq n-1 \) and \( \sigma(j) \).

By combining this result with the Ewens’ sampling formula, we see that if \((a_1, \cdots, a_n)\) is an allelic partition with \( k = a_1 + \cdots + a_n \) alleles, then

\[
\mathbb{P}_{\theta}(A_n = (a_1, \cdots, a_n)|K_n = k) = \frac{\mathbb{P}_{\theta}(A_n = (a_1, \cdots, a_n))}{\mathbb{P}_{\theta}(K_n = k)} = \frac{n!}{|S_n^k|} \prod_{j=1}^{n} \left( \frac{1}{j} \right)^{a_j} \frac{1}{a_j!}.
\]

However, since this conditional probability does not depend on \( \theta \), it follows that

**Theorem 5.** \( K_n \) is a sufficient statistic for \( \theta \).

It follows from this theorem that we need not look beyond \( K_n \) when constructing estimators of \( \theta \) within the context of the infinite alleles model. One such estimator is the maximum-likelihood estimator, which we denote \( \hat{\theta}_{MLE} \), which is the value of \( \theta \in [0, \infty) \) which maximizes the likelihood of the observed number of alleles. Since \( \ln \) is a monotonically increasing function, \( \hat{\theta}_{MLE} \) also maximizes the log-likelihood function, which is equal to

\[
l_n(\theta, k) = \ln(\mathbb{P}_{\theta}(K_n = k)) = \ln(|S_n^k|) + k \ln(\theta) - \sum_{i=1}^{n} \ln(\theta + i - 1).
\]

If we differentiate with respect to \( \theta \) and set the result equal to 0, we obtain the identity,

\[
\frac{k}{\hat{\theta}_{MLE}} - \sum_{i=1}^{n} \frac{1}{\hat{\theta}_{MLE} + i - 1} = 0,
\]

which can be rearranged to give

\[
k = \sum_{i=1}^{n} \frac{\hat{\theta}_{MLE}}{\hat{\theta}_{MLE} + i - 1} = \mathbb{E}_{\hat{\theta}_{MLE}}[K_n].
\]

In other words, the maximum-likelihood estimate of \( \theta \) is the value that makes the expected number of alleles equal to the observed number, i.e., in this case, the MLE is also a method-of-moments estimator.

### 2 Infinite Sites Model

#### 2.1 Motivation

The infinite sites model is similar to the infinite alleles model, but differs in its treatment of mutation. It was introduced by Kimura (1969) and makes the following assumptions:

1. Reproduction occurs according to the Wright-Fisher model.
2. Each chromosome differs from its parent at a Poisson-distributed number of sites with mean $\mu$.

3. Each mutation affects a new site.

As in the previous section, we will continue to investigate a diploid population containing $2N$ chromosomes, although all of our results can be applied to haploid populations by replacing $2N$ by $N$. We will also continue to use $\theta = 4N\mu$ to denote the scaled mutation rate. The name reflects the fact that we can derive this model as the limit of a sequence of $n$-site models by letting $n \to \infty$ while holding the total mutation rate $\mu = n \cdot \nu$ constant. Here $\nu$ is the mutation rate per site. Provided that $n$ is large and that $\nu$ is sufficiently small relative to $N$, which is often the case, the infinite sites model is a good approximation to the more biologically-realistic $n$-sites model. However, because each site can mutate at most once under this model, we need only keep track of two alleles per site, which are conventionally described as the wild type allele and the mutant or derived allele. Another convention found in the literature is to denote the wild type allele by 0 and the derived allele by 1, so that each haplotype segregating in the population can be represented by a binary sequence of 0’s and 1’s.

### 2.2 Segregating sites

Once again our focus will be on the sampling distributions of various summary statistics that can be calculated from a random sample of $n$ chromosomes. Let us say that a site is segregating if our sample contains both mutant and wild type alleles at that position. In this section we will investigate the distribution of the number of segregating sites in a sample of size $n$, which we denote $S_n$. We begin by observing that we can generate samples of chromosomes carrying the correct distribution of mutated sites in two equivalent ways:

1. Simulate the genealogy $T$ using Kingman’s coalescent and then run an independent Poisson process of intensity $\theta/2$ on each branch of $T$. Whenever a mutation occurs along a branch, a new site is mutated which is then inherited by all of the chromosomes descended from the mutated individual.

2. Alternatively, we can simulate the coalescent with mutation, which generates both the genealogy and the mutations concurrently. This process has the same transition rates as the process described in Approach 3 in Section 1.2, i.e., each pair of lineages coalesces at rate 1, while each lineage mutates at rate $\theta/2$. (We continue to measure time in units of $2N$ generations.)

However, because every mutation affects a different site under the infinite sites model, we cannot use either the coalescent with killing or Hoppe’s urn to generate random samples for this model. Rather, every lineage ancestral to a sampled chromosome must be followed all the way back to the MRCA. For this reason, the infinite sites model is more complicated than the infinite alleles model, but we can still use genealogical arguments to derive both the mean and the variance of $S_n$.

**Theorem 6.** Under the infinite sites model, the mean and the variance of $S_n$ are

$$
\begin{align*}
\mathbb{E}[S_n] &= \theta h_n \\
\text{Var}(S_n) &= \theta h_n + \theta^2 g_n,
\end{align*}
$$

where the constants $h_n$ and $g_n$ are defined by

$$
\begin{align*}
h_n &= \sum_{j=1}^{n-1} \frac{1}{j} \\
g_n &= \sum_{j=1}^{n-1} \frac{1}{j^2}.
\end{align*}
$$
Proof. To calculate the mean and the variance of $S_n$, let $\tau_j$ denote the length of time (in units of $2N$ generations) during which there are $j$ lineages in the genealogy of the sample, $\mathcal{T}$, and observe that the total branch length in the genealogy is equal to

$$T_{tot} = \sum_{j=2}^{n} j\tau_j.$$ 

Since the $\tau_j$ are independent exponentially-distributed random variables with parameters $2/j(j-1)$, it follows that the expected total branch length is

$$E[T_{tot}] = \sum_{j=2}^{n} jE[\tau_j] = \sum_{j=2}^{n} \frac{2j}{j(j-1)} = 2h_n,$$

while the variance is

$$\text{Var}(T_{tot}) = \sum_{j=2}^{n} j^2 \cdot \text{Var}(\tau_j) = \sum_{j=2}^{n} \frac{4}{(j-1)^2} = 4g_n.$$

Furthermore, since each mutation on $\mathcal{T}$ generates a new segregating site in the sample, the conditional distribution of $S_n$ given $T_{tot}$ is Poisson, with mean and variance

$$E[S_n|T_{tot}] = \text{Var}(S_n|T_{tot}) = \frac{\theta}{2} T_{tot}.$$ 

The unconditional mean and variance of $S_n$ can then be calculated with the help of the law of total expectation (aka, the tower property) and the law of total variance:

$$E[S_n] = E[E[S_n|T_{tot}]] = \frac{\theta}{2} E[T_{tot}] = \theta h_n,$$

$$\text{Var}(S_n) = E[\text{Var}(S_n|T_{tot})] + \text{Var}(E[S_n|T_{tot}])$$

$$= \frac{\theta}{2} E[T_{tot}] + \left(\frac{\theta}{2}\right)^2 \text{Var}(T_{tot})$$

$$= \theta h_n + \theta^2 g_n.$$ 

Because the coalescent with mutation depends on $N$ and $\mu$ only through the composite parameter $\theta$, it is still the case that the population size and the mutation rate are not identifiable under the infinite sites model. The best that we can hope for is to estimate $\theta$ itself. The preceding theorem suggests an obvious candidate, which is known as Watterson’s estimator

$$\theta_W = S_n/h_n.$$ 

The following result is an immediate consequence of Theorem 6.

**Theorem 7.** Watterson’s estimator is unbiased, i.e., $E[\theta_W] = \theta$, and has variance

$$\text{Var}(\theta_W) = \theta \frac{1}{h_n} + \theta^2 \frac{g_n}{h_n^2}.$$ 

Since $h_n \sim \log(n)$, while $g_n \to \frac{\pi^2}{6}$ as $n \to \infty$, it follows that the variance of Watterson’s estimator decreases to zero, but at a logarithmically slow rate. Furthermore,
Theorem 8. Watterson's estimator is asymptotically normal: for every \( x \in (-\infty, \infty) \),

\[
\lim_{n \to \infty} \mathbb{P}_\theta \left( \frac{\theta_W - \theta}{\sqrt{\text{Var}(\theta_W)}} \leq x \right) = \mathbb{P}(Z \leq x),
\]

where \( Z \) is a standard normal random variable.

Proof. Let \( T \) be the genealogy of the sample and let \( s_j \) be the number of mutations that occurred when there were \( j \) branches in \( T \). Then \( s_2, \ldots, s_n \) are independent and \( S_n = s_2 + \cdots + s_n \), since each mutation on \( T \) creates one of the segregating sites in the sample. To identify the distribution of \( s_j \), recall that the transition probabilities for the jump chain of the coalescent with mutation depend only on the number of branches in the tree and that this number is only changed by coalescent events, but not by mutations. In particular, the probability that the next event is a mutation is \( \theta/(\theta + j - 1) \) when there are \( j \) branches in the genealogy. It follows that \( s_j \) is geometrically-distributed on \( \{0, 1, \cdots\} \), with probabilities

\[
\mathbb{P}_\theta(s_j = k) = \left( \frac{\theta}{\theta + j - 1} \right)^k \frac{j - 1}{\theta + j - 1}
\]

for \( k \geq 0 \). In particular, the variance of \( s_j \) is equal to

\[
\text{Var}(s_j) = \frac{\theta}{j - 1} + \frac{\theta^2}{(j - 1)^2},
\]

which is small relative to \( \text{Var}(S_n) \) when \( n \) is large. Asymptotic normality of \( S_n \) and therefore of \( \theta_W \) then follows from the Feller-Lindeberg central limit theorem.

We can also show that Watterson's estimator is close to optimal in a certain sense.

Definition 2. Suppose that \( \mathcal{P} = \{p(x; \theta) : \theta \in \Theta\} \) is a one-dimensional parametric family of distributions, where \( \Theta \subset \mathbb{R} \), such that the densities \( p(x; \theta) \) depend differentially on \( \theta \). The **Fisher information** of \( \mathcal{P} \) at \( \theta \) is the quantity

\[
\mathcal{I}(\theta) = \mathbb{E}_\theta \left[ \left( \frac{d}{d\theta} l(X; \theta) \right)^2 \right],
\]

where \( l(x; \theta) = \log(p(x; \theta)) \) is the log-likelihood function and \( \mathbb{E}_\theta \) is the expectation with respect to the distribution \( p(\cdot; \theta) \).

The Fisher information for a parametric family measures how strongly the likelihood of the data depends, on average, on the value of the parameter \( \theta \). In particular, when \( \mathcal{I}(\theta) \) is large, the likelihood is strongly dependent on \( \theta \) and so we can expect the data to be very informative about the true value of the parameter. Under fairly mild regularity conditions on the likelihood function, we can also calculate the Fisher information using the following alternative formula

\[
\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[ \frac{d^2}{d\theta^2} l(X; \theta) \right].
\]

Here we will exploit a classical result from the theory of statistics, known as the **Cramér-Rao bound**, which gives a lower bound for the variance of an unbiased estimator of an unknown parameter.
**Theorem 9.** Suppose that \( \hat{\theta} = \hat{\theta}(X) \) is an unbiased estimator of a parameter \( \theta \), where the distribution of the data \( X \) belongs to a one-dimensional parametric family \( P = \{p(x; \theta) : \theta \in \Theta\} \). Then
\[
\text{Var}(\hat{\theta}(X)) \geq \frac{1}{I(\theta)},
\]
when \( X \sim p(\cdot, \theta) \) and \( I(\theta) \) is the Fisher information of \( P \) at \( \theta \).

We will use this result to find a lower bounded for the variance of an unbiased estimator of the scaled mutation rate \( \theta \) under the infinite sites model. In fact, we will assume that we have access to additional information in the form of the number of segregating sites, \( s_j \), created when the genealogy of the sample contained \( j \) branches, for \( j = 2, \cdots, n \). In practice, this information is not available from sequence data and so the lower bound given in the next theorem is conservative in the sense that it may be lower than the bound that can be achieved by an unbiased estimator of \( \theta \) based on the sequence data itself.

**Theorem 10.** Suppose that \( \hat{\theta} = \hat{\theta}(s_2, \cdots, s_n) \) is an unbiased estimator of \( \theta \). Then
\[
\text{Var}(\hat{\theta}) \geq J_n(\theta) = \frac{\theta}{n} - \frac{1}{\sum_{k=1}^{n-1} \frac{1}{k + \theta}}.
\]

**Proof.** Since the \( s_j \) are independent geometrically-distributed random variables with success probabilities \( p_j = (j - 1)/(\theta + j - 1) \), the likelihood function of \( (s_2, \cdots, s_n) \) is
\[
L_n(\theta) = \prod_{j=2}^{n} \left( \frac{\theta}{\theta + j - 1} \right)^{s_j} \frac{j - 1}{\theta + j - 1} = (n - 1)! \theta^{S_n} \prod_{j=2}^{n} (\theta + j - 1)^{-(s_j + 1)},
\]
while the log-likelihood function is
\[
l_n(\theta) = \log(L_n(\theta)) = \log((n - 1)!) + S_n \log(\theta) - \sum_{j=2}^{n} (s_j + 1) \log(\theta + j - 1).
\]
Differentiating twice with respect to \( \theta \) gives
\[
\frac{d^2}{d\theta^2} l_n(\theta) = -\frac{S_n}{\theta^2} + \sum_{j=2}^{n} \frac{s_j + 1}{(\theta + j - 1)^2}.
\]
Since
\[
\mathbb{E}_\theta[s_j + 1] = \frac{1}{p_j} = \frac{\theta + j - 1}{j - 1}
\]
and \( \mathbb{E}_\theta[S_n] = \theta h_n \), the alternative formula for the Fisher information gives
\[
I_n(\theta) = \frac{h_n}{\theta} - \sum_{j=2}^{n} \frac{1}{(j - 1)(\theta + j - 1)}
\]
\[
= \frac{1}{\theta} \sum_{k=1}^{n-1} \left( \frac{1}{k} - \frac{\theta}{k(\theta + k)} \right)
\]
\[
= \frac{1}{\theta} \sum_{k=1}^{n-1} \frac{1}{k + \theta},
\]
i.e., \( I(\theta) = 1/J_n(\theta) \).
Since \( J_n(\theta) \sim \theta/h_n \) as \( n \to \infty \), it follows from Theorem 7 that Watterson’s estimator asymptotically achieves the Cramér-Rao lower bound derived in Theorem 10, i.e.,

\[
\lim_{n \to \infty} \frac{\text{Var}(\theta_W)}{J_n(\theta)} = 1.
\]

In particular, there is no way to improve on the logarithmic rate of decrease of the variance if we restrict ourselves to unbiased estimators.

### 2.3 Pairwise differences

Another summary statistic that is commonly used in DNA sequence analysis is the **mean number of pairwise differences** in a sample of \( n \) chromosomes. This is denoted \( \Delta_n \) (or sometimes \( \theta \pi \)) and is defined by the formula

\[
\Delta_n = \frac{1}{(n/2)} \sum_{i<j} \Delta_{ij}
\]

where \( \Delta_{ij} \) is the number of pairwise differences between sequence \( i \) and sequence \( j \), i.e., the number of sites at which they carry different nucleotides. When \( n = 2 \), this is just the number of segregating sites in the sample and so \( S_2 = \Delta_2 \), but for \( n > 2 \) we will generally have \( \Delta_n < S_n \).

Perhaps the most important difference between these two statistics is that whereas \( \Delta_n \) depends on the frequencies of the two nucleotides found at each segregating site, \( S_n \) gives equal weight to every segregating site. Our next theorem asserts that \( \Delta_n \) is an unbiased estimator of \( \theta \).

**Theorem 11.** Under the infinite sites model, \( \mathbb{E}_\theta[\Delta_n] = \theta \).

**Proof.** Provided that the sampled chromosomes have been assigned random labels, the expected value \( \mathbb{E}_\theta[\Delta_{ij}] \) does not depend on \( i \) and \( j \) and so \( \mathbb{E}_\theta[\Delta_n] = \mathbb{E}_\theta[\Delta_{ij}] \) for every pair \( i \neq j \). So fix \( i \neq j \) and consider the coalescent process for these two chromosomes under the infinite sites model. Since mutations occur at rate \( 2 \times \theta/2 = \theta \) (each lineage mutates at rate \( \theta/2 \)) and coalescence occurs at rate 1, the transition probabilities of the embedded jump chain are:

\[
\mathbb{P}_\theta(\text{mutation}) = \frac{\theta}{\theta+1} \quad \text{and} \quad \mathbb{P}_\theta(\text{coalescence}) = \frac{1}{\theta+1}.
\]

Furthermore, since each mutation contributes a new pairwise difference, but does not change the number of lineages, \( \Delta_{ij} \) is equal to the number of mutations that occur along either lineage before they coalesce. This number is geometrically-distributed on \( \{0, 1, \cdots\} \) with success probability \( p = 1/(\theta + 1) \) and so

\[
\mathbb{P}_\theta(\Delta_{ij} = k) = \left( \frac{\theta}{\theta+1} \right)^k \frac{1}{\theta+1}
\]

for \( k \geq 0 \). In particular, the mean of this distribution is

\[
\mathbb{E}_\theta[\Delta_{ij}] = \frac{1}{p} - 1 = \theta,
\]

which completes the proof. \( \square \)

Given a sample of \( n \) chromosomes, let \( \eta_k, k = 1, \cdots, n-1 \) be the number of sites at which exactly \( k \) chromosomes carry the derived nucleotide and observe that the nucleotide diversity at each of these sites is equal to \( k(n-k)/(n/2) \). Since the mean number of pairwise differences is equal to
the sum of the nucleotide diversities over all of the segregating sites, it follows that $\Delta_n$ is also equal to the following sum

$$\Delta_n = \sum_{k=1}^{n-1} \left( \frac{2}{n(n-1)} k(n-k) \right) \eta_k.$$ 

In fact, there is an entire family of estimators of $\theta$ which have the following general form:

$$\hat{\theta} = \sum_{k=1}^{n-1} c_{n,k} \eta_k,$$

where $c_{n,k}$ are constants. For example, since $S_n = \eta_1 + \cdots + \eta_{n-1}$, it follows that Watterson’s estimator $\theta_W = S_n/h_n$ is also of this form, with $c_{n,k} = h_n^{-1}$. The following theorem asserts that the variance of any such estimator is a quadratic function of $\theta$.

**Theorem 12.** Under the infinite sites model, the variance of the estimator $\hat{\theta} = \sum_{k=1}^{n-1} c_{n,k} \eta_k$ has the form

$$\text{Var}_\theta(\hat{\theta}) = a_n \theta + b_n \theta^2,$$

where $a_n$ and $b_n$ are constants that depend on the $c_{n,k}$.

**Proof.** Let $T$ be the random genealogy of the sample of $n$ chromosomes and let $L_k$, $k = 1, \cdots, n-1$ be the total length of the branches that have exactly $k$ descendants in the sample. For example, $L_1$ will be the total length of the branches that directly terminate in the leaves of the tree. Notice that under the infinite sites model each mutation occurring on such a branch will be inherited by exactly $k$ chromosomes in the sample. Consequently, since mutations along $T$ follow an independent Poisson process of intensity $\theta/2$ on each branch, it follows that $\eta_1, \cdots, \eta_{n-1}$ are conditionally independent given $\mathcal{L} = (L_1, \cdots, L_{n-1})$ and that $\eta_k$ is Poisson-distributed with mean and variance equal to $\theta L_k/2$. In particular, the conditional mean and variance of $\hat{\theta}$ are given by

$$\mathbb{E}_\theta(\hat{\theta}|\mathcal{L}) = \frac{\theta}{2} \sum_{k=1}^{n-1} c_{n,k} L_k \quad \text{Var}_\theta(\hat{\theta}|\mathcal{L}) = \frac{\theta}{2} \sum_{k=1}^{n-1} c_{n,k}^2 L_k.$$

We can then use the law of total expectations and the law of total variance to calculate

$$\mathbb{E}_\theta(\hat{\theta}) = \frac{\theta}{2} \sum_{k=1}^{n-1} c_{n,k} \mathbb{E}[L_k]$$

$$\text{Var}_\theta(\hat{\theta}) = \mathbb{E} \left[ \text{Var}_\theta(\hat{\theta}|\mathcal{L}) \right] + \text{Var} \left( \mathbb{E}_\theta(\hat{\theta}|\mathcal{L}) \right)$$

$$= \left( \frac{1}{2} \sum_{k=1}^{n-1} c_{n,k}^2 \mathbb{E}[L_k] \right) \theta + \left( \frac{1}{4} \sum_{j,k=1}^{n-1} c_{n,j} c_{n,k} \text{Cov}(L_j, L_k) \right) \theta^2$$

$$= a_n \theta + b_n \theta^2,$$

which completes the proof. \qed

This result is valid for a much larger class of neutral population genetical models that satisfy the infinite sites mutation assumption, e.g., we can modify the Wright-Fisher model to allow for population growth or structure. What will change are the coefficients $a_n$ and $b_n$ since these depend on the distribution of branch lengths in the genealogy of a random sample through the means and covariances of the random variables $L_1, \cdots, L_{n-1}$. In particular, it can be shown (Tajima, 1983) that

$$\text{Var}_\theta(\Delta_n) = \frac{n+1}{3(n-1)} \theta + \frac{2(n^2 + n + 3)}{9n(n-1)} \theta^2,$$
which has the following large sample limit

\[
\lim_{n \to \infty} \text{Var}_\theta(\Delta_n) = \frac{1}{3} \theta + \frac{2}{9} \theta^2.
\]

Thus, unlike Watterson’s estimator, the variance of \( \Delta_n \) does not tend to zero as the sample size increases to infinity and so the latter estimator, although unbiased, cannot be made arbitrarily precise by increasing the sample size. However, as we will see later, \( \Delta_n \) is still useful in the context of tests of neutrality.