

Estimation and Hypothesis Testing

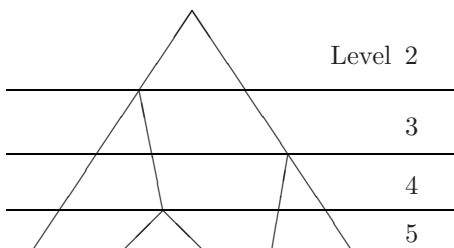
“It is easy to lie with statistics. It is hard to tell the truth without it.” Andrejs Dunkels

2.1 Site frequency spectrum covariance

Assume the Wright-Fisher model with infinitely many sites, and let η_i be the number of sites where the mutant (or derived) allele has frequency i in a sample of size n . In this section, we will describe work of Fu (1995), which allows us to compute $\text{var}(\eta_i)$ and $\text{cov}(\eta_i, \eta_j)$. We begin with the new proof of the result for the mean given in Theorem 1.33.

Theorem 2.1. $E\eta_i = \theta/i$.

Proof. We say that a time t is at level k if there are k sample lineages in the coalescent at that time. The key to the proof is to break things down according to the level at which mutations occur.



Let L_m be the total length of branches with m descendants. Let J_ℓ^k be the number of sampled individuals that are descendants of edge ℓ at level k .

Theorem 1.6 implies that (J_1^k, \dots, J_k^k) is uniformly distributed over the vectors of k positive integers that add up to n . Recalling that the number of such vectors is $\binom{n-1}{k-1}$, it follows that for $1 \leq i \leq n - k + 1$

$$P(J_\ell^k = i) = \binom{n-i-1}{k-2} / \binom{n-1}{k-1} \quad (2.1)$$

since the numerator gives the number of ways of breaking the remaining $n - i$ individuals into $k - 1$ nonempty groups.

Since level k lasts for an amount of time with mean $2/k(k-1)$ and there are k edges on level k ,

$$EL_i = \sum_{k=2}^n \frac{2}{k(k-1)} \cdot k \cdot \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \quad (2.2)$$

Since mutations occur on the tree at rate $\theta/2$, it suffices to show $EL_i = 2/i$. To evaluate the sum, we need the following identity:

$$\frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \frac{1}{k-1} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} \frac{1}{i} \quad (2.3)$$

Skipping the proof for the moment and using this in (2.2) gives

$$EL_i = \frac{2}{i} \sum_{k=2}^n \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} = \frac{2}{i}$$

To see the last equality, note that if we are going to pick i things out of $n - 1$ and the index of the first one chosen is $k - 1$, with $2 \leq k \leq n$, then we must choose $i - 1$ from the last $(n - 1) - (k - 1)$ items.

The last detail is to prove (2.3). Recalling the definition of the binomial coefficients,

$$\frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \frac{1}{k-1} = \frac{1}{k-1} \frac{(n-i-1)!}{(k-2)!(n-i-k+1)!} \cdot \frac{(k-1)!(n-k)!}{(n-1)!}$$

Cancelling the $(k-1)!$, swapping the positions of $(n-i-1)!$ and $(n-k)!$, and then multiplying by $i!/(i-1)!i$, the above becomes

$$= \frac{(n-k)!}{i(i-1)!(n-k-(i-1))!} \cdot \frac{i!(n-i-1)!}{(n-1)!} = \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} \frac{1}{i}$$

and the proof is complete. \square

To state Fu's (1995) result for the covariances, we recall $h_n = \sum_{i=1}^{n-1} 1/i$ and let

$$\beta_n(i) = \frac{2n}{(n-i+1)(n-i)}(h_{n+1} - h_i) - \frac{2}{n-i}$$

Theorem 2.2. $\text{var}(\eta_i) = \theta/i + \theta^2\sigma_{ii}$ and for $i \neq j$, $\text{cov}(\eta_i, \eta_j) = \sigma_{ij}\theta^2$. The diagonal entries σ_{ii} are given by

$$\begin{aligned} \beta_n(i+1) & \quad i < n/2 \\ 2\frac{h_n - h_i}{n-i} - \frac{1}{i^2} & \quad i = n/2 \\ \beta_n(i) - \frac{1}{i^2} & \quad i > n/2 \end{aligned} \quad (2.4)$$

while σ_{ij} with $i > j$ are given by

$$\begin{aligned} \frac{\beta_n(i+1) - \beta_n(i)}{2} & \quad i+j < n \\ \frac{h_n - h_i}{n-i} + \frac{h_n - h_j}{n-j} - \frac{\beta_n(i) + \beta_n(j+1)}{2} - \frac{1}{ij} & \quad i+j = n \\ \frac{\beta_n(j) - \beta_n(j+1)}{2} - \frac{1}{ij} & \quad i+j > n \end{aligned} \quad (2.5)$$

To help understand the formulas, we begin by computing σ_{ij} in the special case $n = 8$.

	$j = 1$	2	3	4	5	6	7
$i = 1$	0.3211	-0.0358	-0.0210	-0.0141	-0.0103	-0.0079	0.1384
2	-0.0358	0.2495	-0.0210	-0.0141	-0.0103	0.1328	-0.0356
3	-0.0210	-0.0210	0.2076	-0.0141	0.1283	-0.0346	-0.0356
4	-0.0141	-0.0141	-0.0141	0.3173	-0.0359	-0.0275	-0.0267
5	-0.0103	-0.0103	0.1283	-0.0359	0.1394	-0.0230	-0.0216
6	-0.0079	0.1328	-0.0346	-0.0275	-0.0230	0.1310	-0.0183
7	0.1384	-0.0356	-0.0267	-0.0216	-0.0183	-0.0159	0.1224

The numbers on the diagonal must be positive since $\text{var}(\eta_i) > 0$. All of the off-diagonal elements are negative numbers, except for the σ_{ij} with $i+j = n$. Intuitively, these are positive due the fact that the first split in the tree may have i lineages on the left and $n-i$ on the right, and this event increases both η_i and η_{n-i} . The negative off-diagonal elements are small, but since there are $O(n^2)$ of them, their sum is significant.

The next two figures give the values of the covariance matrix when $N = 25$. The first gives the values on the diagonal $i = j$ and the anti-diagonal $i+j = 25$. Since 25 is odd, these do not intersect. The bump in the middle of the graph of the diagonal covariances may look odd, but when $i \geq 13$, there cannot be two edges in the tree that produce these mutations. To better see the off-diagonal entries in the covariance matrix in the second figure, we have plotted $-\sigma_{ij}$ and inserted 0's on the two diagonals. The largest value is at σ_{12} . Note the jump in the size of the correlations when we enter the region $i+j > 25$.

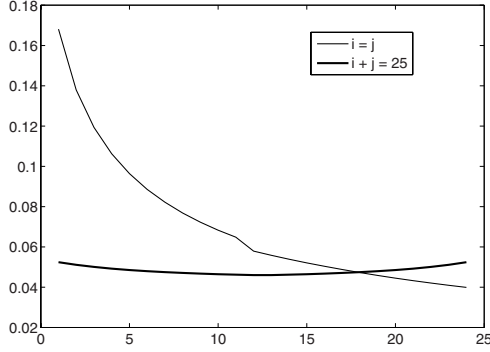


Fig. 2.1. Diagonals of the covariance matrix when $N = 25$.

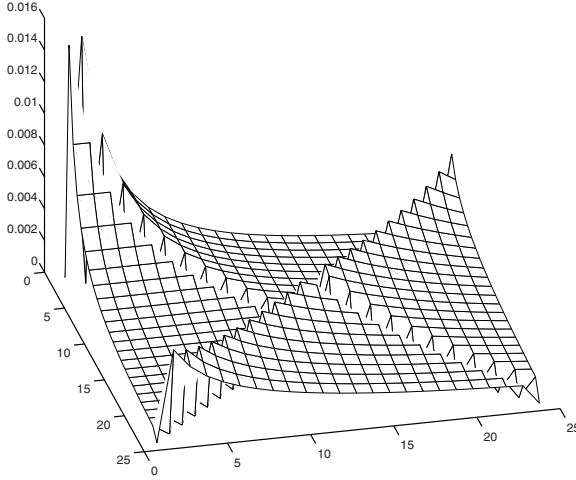


Fig. 2.2. Off-diagonal entries (times -1) of the covariance matrix when $N = 25$.

Proof. Let $\nu_{\ell,k}$ = the number of mutations that occur on edge ℓ at level k . To compute variances and covariances, we use the identity

$$\eta_i = \sum_{k=2}^n \sum_{\ell=1}^k 1_{(J_{\ell}^k=i)} \nu_{\ell,k}$$

where the indicator function $1_{(J_{\ell}^k=i)}$ is equal to 1 if $J_{\ell}^k = i$ and 0 otherwise. Multiplying two copies of the sum and sorting the terms, we conclude that

$$\begin{aligned}
E(\eta_i \eta_j) &= 1_{(i=j)} \sum_{k=2}^n k P(J_1^k = i) E\nu_{1,k}^2 \\
&+ \sum_{k=2}^n k(k-1) P(J_1^k = i, J_2^k = j) E(\nu_{1,k} \nu_{2,k}) \\
&+ \sum_{k \neq h} k h P(J_1^k = i, J_1^h = j) E(\nu_{1,k} \nu_{1,h})
\end{aligned} \tag{2.6}$$

To evaluate the last expression, we have to compute (i) the expected values of the $\nu_{\ell,k}$, and (ii) the joint distribution of the J_ℓ^k . Readers who lose interest in the details can skip to the result given in Theorem 2.2.

The expected values of the $\nu_{\ell,k}$ are easy to compute. Let t_k be the amount of time at level k . Conditioning on the value of t_k ,

$$(\nu_{\ell,k} | t_k = t) = \text{Poisson}(\theta t/2)$$

Since t_k is exponential with mean $2/k(k-1)$,

$$E\nu_{\ell,k} = \frac{\theta}{k(k-1)} \tag{2.7}$$

Reasoning as in the derivation of (1.23), recalling $\text{Poisson}(\theta t/2)$ has variance $\theta t/2$, and using $\text{var}(t_k) = 4/(k^2(k-1)^2)$, we have

$$\begin{aligned}
\text{var}(\nu_{\ell,k}) &= E(\text{var}(\nu_{\ell,k} | t_k)) + \text{var}(E(\nu_{\ell,k} | t_k)) \\
&= \frac{\theta}{2} E(t_k) + \frac{\theta^2}{4} \text{var}(t_k) \\
&= \frac{\theta}{k(k-1)} + \frac{\theta^2}{k^2(k-1)^2}
\end{aligned}$$

Adding the square of the mean we have

$$E\nu_{\ell,k}^2 = \frac{\theta}{k(k-1)} + \frac{2\theta^2}{k^2(k-1)^2} \tag{2.8}$$

Conditioning on the values of t_k and t_h , and using $E(t_k^2) = 8/(k^2(k-1)^2)$,

$$\text{for } \ell \neq m \quad E(\nu_{\ell,k} \nu_{m,k}) = \frac{\theta^2}{4} E t_k^2 = \frac{2\theta^2}{k^2(k-1)^2} \tag{2.9}$$

$$\text{for } k \neq h \quad E(\nu_{\ell,k} \nu_{m,h}) = \frac{\theta^2}{4} E t_k E t_h = \frac{\theta^2}{k(k-1)h(h-1)} \tag{2.10}$$

The joint probabilities for pairs of edges are more difficult to compute. The easiest situation occurs when both edges are on the same level. When $k = 2$, Tajima's result, Theorem 1.6, tells us that

$$P(J_1^2 = i, J_2^2 = n - i) = \frac{1}{n-1} \quad (2.11)$$

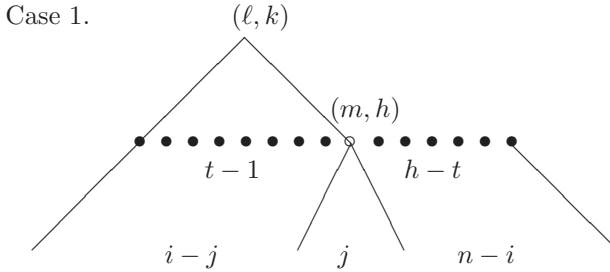
Extending the reasoning for (2.1) shows that for $k \geq 3$ and $i + j < n$

$$P(J_\ell^k = i, J_m^k = j) = \frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}} \quad (2.12)$$

since the remaining $n - (i + j)$ sampled individuals are divided between the other $k - 2$ lineages. Similar reasoning shows that if $2 \leq t \leq k - 1$ and $j < i < n$,

$$P\left(J_1^k = j, \sum_{\ell=2}^t J_\ell^k = i - j\right) = \frac{\binom{i-j-1}{t-2} \binom{n-i-1}{k-t-1}}{\binom{n-1}{k-1}} \quad (2.13)$$

Suppose now that $k < h$ and let $D_{\ell,k}^h$ be the number of descendants on level h of edge (ℓ, k) , i.e., edge ℓ on level k . Writing $(m, h) < (\ell, k)$ for (m, h) is a descendant of (ℓ, k) , there are two cases to consider: $(m, h) < (\ell, k)$ and $(m, h) \not< (\ell, k)$. In each case there is the general situation and a degenerate subcase in which one group of vertices is empty.



Here we have not drawn the tree, but have indicated the sizes of the various sets of descendants. Recalling that edges on each level are randomly labeled, we see that if $j < i < n$ and $t \geq 2$,

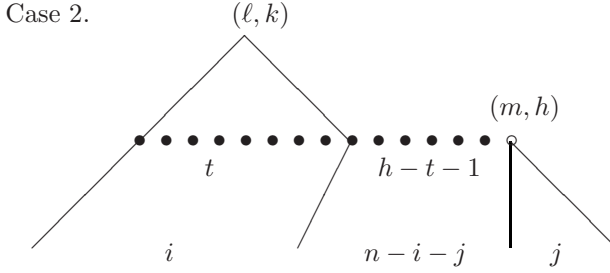
$$\begin{aligned} P(D_{\ell,k}^h = t, (m, h) < (\ell, k), J_\ell^k = i, J_m^h = j) \\ = \frac{\binom{h-t-1}{k-2}}{\binom{h-1}{k-1}} \cdot \frac{t}{h} \cdot \frac{\binom{i-j-1}{t-2} \binom{n-i-1}{h-t-1}}{\binom{n-1}{h-1}} \end{aligned} \quad (2.14)$$

The first factor is $P(D_{\ell,k}^h = t)$, computed from (2.1). The second is $P((m, h) < (\ell, k) | D_{\ell,k}^h = t)$. When the first two events occur, we need $J_m^h = j$ and the other $t - 1$ descendants of (ℓ, k) on level h to have a total of $i - j$ descendants, so the desired result follows from (2.13).

If $i = j$, then we must have $t = 1$ and

$$\begin{aligned}
P(D_{\ell,k}^h = 1, (m, h) < (\ell, k), J_\ell^k = i, J_m^h = i) \\
= \frac{\binom{h-2}{k-2}}{\binom{h-1}{k-1}} \cdot \frac{1}{h} \cdot \frac{\binom{n-i-1}{h-2}}{\binom{n-1}{h-1}}
\end{aligned} \tag{2.15}$$

If one is willing to adopt the somewhat strange convention that $\binom{-1}{-1} = 1$, this can be obtained from the previous formula by setting $t = 1$ and $i - j = 0$.



Reasoning similar to Case 1 shows that if $t \leq h - 2$ and $i + j < n$,

$$\begin{aligned}
P(D_{\ell,k}^h = t, (m, h) \not< (\ell, k), J_\ell^k = i, J_m^h = j) \\
= \frac{\binom{h-t-1}{k-2}}{\binom{h-1}{k-1}} \cdot \frac{h-t}{h} \cdot \frac{\binom{i-1}{t-1} \binom{n-(i+j)-1}{h-t-2}}{\binom{n-1}{h-1}}
\end{aligned} \tag{2.16}$$

If $i + j = n$, then we must have $t = h - 1$ and $k = 2$. (We may have $h > 2$.)

$$\begin{aligned}
P(D_{\ell,k}^h = h - 1, (m, h) \not< (\ell, k), J_\ell^k = i, J_m^h = n - i) \\
= \frac{1}{h-1} \cdot \frac{1}{h} \cdot \frac{\binom{i-1}{h-2}}{\binom{n-1}{h-1}}
\end{aligned} \tag{2.17}$$

Again, if $\binom{-1}{-1} = 1$, this can be obtained from the previous formula by setting $t = h - 1$ and $n - (i + j) = 0$.

At this point we have everything we need to compute the variances and covariances. The expressions in (2.14)–(2.17) are complicated and two of them must be summed over t . Remarkably, Fu (1995) was able to obtain the given analytical formulas for the quantities of interest. Many details need to end up with these results, so we refer the reader to the original article for details.

2.2 Estimates of θ

Let η_i be the number of sites where the mutant allele is present i times in a sample of size n . There are a number of ways of estimating $\theta = 4Nu$, where u

is the mutation rate for the locus, using linear functions of the site frequency spectrum, that is, estimators of the form

$$\hat{\theta} = \sum_{i=1}^{n-1} c_{n,i} \eta_i.$$

Fu and Li (1993). $\theta_{FL} = \eta_1$.

Watterson (1975). $\theta_W = h_n^{-1} \sum_{i=1}^{n-1} \eta_i$, where $h_n = \sum_{i=1}^{n-1} 1/i$.

Zeng, Fu, Shi, and Wu (2006). $\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i \eta_i$

Tajima (1983). $\theta_\pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i) \eta_i$.

Fay and Wu (2000). $\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \eta_i = 2\theta_L - \theta_\pi$.

To check that these are unbiased estimators, we note that $E\hat{\theta} = \sum_{i=1}^{n-1} c_{n,i} \theta/i$. The fact that $E\hat{\theta} = \theta$ in the first three cases is easy to see. For the fourth and fifth, we note that

$$\sum_{i=1}^{n-1} i = \frac{n(n-1)}{2} = \sum_{i=1}^{n-1} (n-i)$$

It is much more difficult to compute the variances. Using $g_n = \sum_{i=1}^{n-1} 1/i^2$, we can write the answers as

$$\text{var}(\theta_W) = \frac{\theta}{h_n} + \frac{g_n}{h_n^2} \theta^2 \quad (2.18)$$

$$\text{var}(\eta_1) = \theta + 2 \frac{nh_n - 2(n-1)}{(n-1)(n-2)} \theta^2 \quad (2.19)$$

$$\text{var}(\theta_\pi) = \frac{n+1}{3(n-1)} \theta + \frac{2(n^2 + n + 3)}{9n(n-1)} \theta^2 \quad (2.20)$$

$$\text{var}(\theta_L) = \frac{n}{2(n-1)} \theta + \left[2 \left(\frac{n}{n-1} \right)^2 (g_{n+1} - 1) - 1 \right] \theta^2 \quad (2.21)$$

$$\text{var}(\theta_H) = \theta + \frac{2[36n^2(2n+1)g_{n+1} - 116n^3 + 9n^2 + 2n - 3]}{9n(n-1)^2} \theta^2 \quad (2.22)$$

We have seen the first and third results in (1.22) and (1.30). The second is due to Fu and Li (1993). The fourth and fifth are from Zeng, Fu, Shi, and Wu (2006). Note that, as we proved in Theorem 1.27, each variance has the form $a_n \theta + b_n \theta^2$. The term with θ is the mutational variance due to the placement of mutations on the tree, while the term with θ^2 is the evolutionary variance due to fluctuations in the shape of the tree.

It is easy to visually compare the terms with θ . In the case of θ_W , $\theta/h_n \rightarrow 0$. η_1 and θ_H have θ , while for θ_π and θ_L the terms are $\approx \theta/3$ and $\approx \theta/2$

respectively. It is hard to understand the relationship between the variances by looking at the formulas. In the next table, we have evaluated the coefficient of θ^2 for the indicated values of n . The limits as $n \rightarrow \infty$ are 0, 0, $2/9$, 0.289863, and 0.541123.

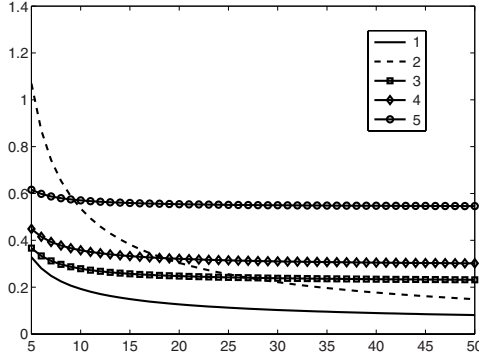


Fig. 2.3. Coefficient of θ^2 in the variance of 1. θ_W , 2. η_1 , 3. θ_π , 4. θ_L , and 5. θ_H .

In addition to formulas for the variances, we will need the following covariances:

$$\text{cov}(\eta_k, S_n) = \frac{\theta}{k} + \frac{h_n - h_k}{n - k} \theta^2 \quad (2.23)$$

$$\text{cov}(\theta_L, S_n) = \theta + \frac{ng_n - (n-1)}{(n-1)} \theta^2 \quad (2.24)$$

$$\text{cov}(\theta_\pi, S_n) = \theta + \left(\frac{n+2}{2n} \right) \theta^2 \quad (2.25)$$

$$\text{cov}(\theta_L, \theta_\pi) = \frac{n+1}{3(n-1)} \theta + \frac{7n^2 + 3n - 2 - 4n(n+1)g_{n+1}}{2(n-1)^2} \quad (2.26)$$

The first for $k = 1$ is formula (25) from Fu and Li (1993). We prove the more general result here. The third is (25) from Tajima (1989). The other two are in (A1) of Zeng, Fu, Shi, and Wu (2006).

Where do these formulas come from?

In principle, the problem is solved by Fu's result for the site frequency spectrum covariance. Given $X_j = \sum_{i=1}^{n-1} c_{n,i}^j \eta_i$, $\text{cov}(X_1, X_2) = a_n \theta + b_n \theta^2$, where

$$a_n = \sum_{i=1}^{n-1} c_{n,i}^1 c_{n,i}^2 / i \quad (2.27)$$

$$b_n = \sum_{i,j} c_{n,i}^1 \sigma_{ij} c_{n,j}^2 \quad (2.28)$$

Of course, $\text{var}(X_1) = \text{cov}(X_1, X_1)$.

If one wants numerical values for a fixed n , then (2.27), (2.28), and Theorem 2.2 lead easily to the desired answer. It is fairly straightforward to use (2.27) to algebraically compute the coefficients a_n , but the b_n are considerably more difficult. To illustrate this point, we will compute $\text{cov}(\eta_k, S_n)$, which is the simplest of these tasks.

Proof of 2.23. In this case, $c_{n,i}^1 = 1$ if $i = k$ and $c_{n,i}^2 = 1$ for all i , so $a_n = 1/k$. Using (2.28) and Theorem 2.2,

$$\begin{aligned} b_n &= \sum_j \sigma_{kj} = \beta_n(k) + \frac{\beta_n(1) - \beta_n(k)}{2} \\ &\quad + (n-k-1) \frac{\beta_n(k) - \beta_n(k+1)}{2} - (n-k-1) \frac{\beta_n(k) - \beta_n(k+1)}{2} \\ &\quad - \frac{\beta_n(1) - \beta_n(n-k)}{2} - \frac{\beta_n(i) + \beta_n(n-k)}{2} \\ &\quad + \frac{h_n - h_k}{n-k} + \frac{h_n - h_{n-k}}{k} - \frac{1}{k} \sum_{j=n-k}^{n-1} \frac{1}{j} \end{aligned}$$

The terms involving β_n add up to 0, and the last two terms cancel, leaving us with

$$b_n = \sum_{j=1}^{n-1} \sigma_{kj} = \frac{h_n - h_k}{n-k} \quad (2.29)$$

which proves (2.23). \square

Using (2.23), one can, with some effort, compute $\text{var}(S_n)$, $\text{cov}(\theta_L, S_n)$, and $\text{cov}(\theta_\pi, S_n)$, since they are all equal to

$$\sum_{k=1}^{n-1} c_{n,k}^1 \text{cov}(\eta_k, S_n)$$

However, in the other cases, the terms involving β_n do not cancel, and we have not been able to derive the other formulas given above from Theorem 2.2. On the other hand, the results in this section allow one to easily compute the variances and covariances numerically, so perhaps such tedious algebraic manipulations are obsolete.

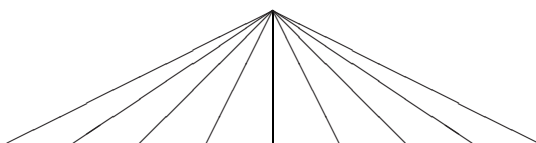
2.3 Hypothesis testing overview

One of the most important questions we face is:

Is the observed DNA sequence data consistent with neutral evolution in a homogeneously mixing population of constant size?

As the last sentence indicates, there are many assumptions that can be violated. Much of the rest of the book is devoted to investigating consequences of alternatives to this null hypothesis. In Chapters 4 and 5, we study population size changes and population structure. In Chapters 6 and 7, we study various types of fitness differences between alleles. Most of the alternative hypotheses can be grouped into two categories.

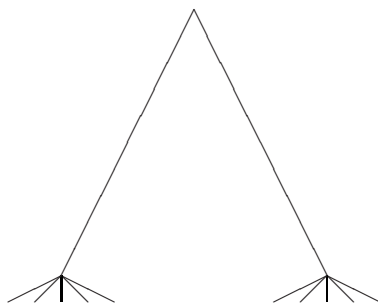
A. Those that tend to make a *star-shaped* genealogy:



Examples of this are:

- *Population bottleneck.* If, as we work backwards in time, there is a sudden decrease in the population size, then the coalescence rate will become large. Situations that can cause this are the founding of a new population by a small number of migrants or, what is essentially the same, improving a crop species by choosing a few individuals with desirable properties.
- *Selective sweep.* This term refers to the appearance of a favorable mutation that rises in frequency until it takes over the population. In the absence of recombination, this is a severe population bottleneck because the entire population will trace its ancestry to the individual with the favorable mutation.
- *Population growth.* In the human population, which has experienced a period of exponential growth, then the coalescence rate will be initially small, and the genealogical tree will have tips that are longer than usual.

B. At the other extreme from a star-shaped genealogy is a *chicken legs genealogy*. There are two long, skinny legs with feet on the ends.



Examples of this are:

- *Population subdivision.* Imagine two isolated populations that exchange migrants infrequently. If we sample 10 individuals from each population, then the two subsamples will coalesce as usual and then eventually the their two ancestral lineages will coalesce.
- *Balancing selection.* This term refers to a situation in which the fitness of heterozygotes Aa is larger than that of homozygotes AA and aa . In this case, the population will settle into a equilibrium in which the two alleles A and a are each present with fixed frequencies. As we will see in Section 6.2, this is essentially a two-population model, with migration between chromosomes with A and chromosomes with a being caused by mutation or recombination.

Notice that in each case there are different explanations that produce the same effect. Thus, one of the problems we will face in hypothesis testing is that if we reject neutral evolution in a homogeneously mixing population of constant size, it will be difficult to say if this is due to natural selection or to demographic factors such as population structure or population size changes.

This is a serious problem for the difference statistics considered in the next section because the tests are performed by comparing the observation to the distribution of the statistic under the neutral model. The HKA test discussed in Section 2.5 avoids this problem by comparing patterns of variability at two regions in the same individuals. The McDonald-Kreitman test discussed in Section 2.6 compares the ratio of nonsynonymous to synonymous polymorphisms within species to the ratio of nonsynonymous to synonymous fixed differences between species, which should not be affected by the shape of the tree.

Here, we will content ourselves to describe the mechanics of the tests and give a few examples to illustrate their use. For more on the issues involved in the use of these and other tests, there are a number of excellent survey

articles: Kreitman (2000), Nielsen (2001), Fay and Wu (2001), Bamshad and Wooding (2003), Nielsen (2005), and Sabeti et al. (2006).

As the reader will see from these articles, there are many tests that we have not discussed. An important omission is the method of demonstrating the presence of positive selection by comparing the number of nonsynonymous mutations per nonsynonymous site (d_N) to the number of nonsynonymous mutations per synonymous site (d_S). Hughes and Nei (1988) showed that $\omega = d_N/d_S > 1$ for the antigen binding cleft of the Major Histocompatibility Complex. A statistical framework for making inferences regarding d_N and d_S was developed by Goldman and Yang (1994) and Muse and Gaut (1994). In this framework, the evolution of a gene is modeled as a continuous-time Markov chain with state space the 61 possible non-stop codons.

In general, testing $\omega < 1$ for an entire gene is a very conservative test of neutrality. Purifying selection often acts on large parts of genes to preserve their function. To address this, Nielsen and Yang (1998) developed a model in which there are three categories of sites: invariable sites ($\omega = 0$), neutral sites ($\omega = 1$), and positively selected sites ($\omega > 1$). Later, Yang et al. (2000) replaced the neutral class by constrained sites that have a distribution of ω values in $(0, 1)$. This test, which is implemented in the computer program PAML, has been used to provide evidence of positive selection in a number of cases; see Nielsen (2001) for some examples.

2.4 Difference statistics

Given two unbiased estimators of θ , we can subtract them to get a random variable with mean 0 that can be used for testing whether the data is consistent with our model of neutral mutations in a homogeneously mixing population of constant size.

2.4.1 Tajima's D

Tajima (1989) was the first to do this, taking the difference $d = \theta_W - \theta_\pi$. We have computed that

$$\begin{aligned} \text{var}(S_n) &= a_1\theta + a_2\theta^2 \quad \text{where} \quad a_1 = \sum_{i=1}^{n-1} 1/i \quad a_2 = \sum_{i=1}^{n-1} 1/i^2 \\ \text{var}(\theta_\pi) &= b_1\theta + b_2\theta^2 \quad \text{where} \quad b_1 = \frac{n+1}{3(n-1)} \quad b_2 = \frac{2(n^2+n+3)}{9n(n-1)} \end{aligned}$$

To compute $\text{var}(d)$, we need (2.25):

$$\text{cov}(S_n, \theta_\pi) = \theta + \left(\frac{1}{2} + \frac{1}{n}\right)\theta^2$$

Recalling $\theta_W = S_n/a_1$, we have

$$\text{var}(d) = c_1\theta + c_2\theta^2 \quad \text{where} \quad c_1 = b_1 - \frac{1}{a_1}, \quad c_2 = b_2 - \frac{n+2}{a_1n} + \frac{a_2}{a_1^2}$$

To finish the definition of Tajima's statistic, we need estimators of θ and θ^2 , so we can construct an estimator of $\text{var}(d)$. For the first we use $\hat{\theta} = S_n/a_1$. For the second we note that

$$E(S_n^2) - ES_n = \text{var}(S_n) + E(S_n)^2 - ES_n = (a_1^2 + a_2)\theta^2$$

so our estimate of $\text{var}(d)$ is $\hat{v}(d) = e_1S_n + e_2S_n(S_n - 1)$, where $e_1 = c_1/a_1$ and $e_2 = c_2/(a_1^2 + a_2)$ and Tajima's test statistic is

$$D_T = \frac{\theta_\pi - \theta_W}{\sqrt{e_1S_n + e_2S_n(S_n - 1)}} \quad (2.30)$$

The smallest value of D_T , call it u , occurs when the minor allele has frequency 1 at each segregating site. This happens when there is a *star-shaped* genealogy. In this case,

$$\theta_\pi - \theta_W = \left[\binom{n}{2}^{-1} (n-1) - \frac{1}{h_n} \right] S_n$$

If S_n is large, we have

$$u \approx \left(\frac{2}{n} - \frac{1}{h_n} \right) / \sqrt{e_2}$$

The largest value of D_T , call it v , occurs where the split between the two nucleotides at each site is as even as it can be. If n is even,

$$\theta_\pi - \theta_W = \left[\binom{n}{2}^{-1} (n/2)^2 - \frac{1}{h_n} \right] S_n$$

If S_n is large, we have

$$v \approx \left(\frac{1}{2} - \frac{1}{h_n} \right) / \sqrt{e_2}$$

Tajima argued on the basis of simulations that the density function of D_T is approximately a generalized beta distribution with range $[u, v]$:

$$f(D) = \frac{\Gamma(\alpha + \beta)(v - D)^{\alpha-1}(D - u)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)(v - u)^{\alpha+\beta-1}}$$

where α and β are chosen to make the mean 0 and the variance 1:

$$\alpha = -\frac{(1 + uv)v}{v - u} \quad \beta = \frac{(1 + uv)u}{v - u}$$

Table 2 in Tajima's paper gives 90%, 95%, 99%, and 99.9% confidence intervals for a variety of sample sizes.

Example 2.1. Aquadro and Greenberg (1983). In Section 1.4, we computed $\theta_W = 17.959184$ and $\theta_\pi = 14.857143$. Using a small computer program, one can compute

$$\begin{array}{ll} a_1 = 2.450000 & a_2 = 1.491389 \\ b_1 = 0.444444 & b_2 = 0.312169 \\ c_1 = 0.036281 & c_2 = 0.035849 \\ e_1 = 0.014809 & e_2 = 0.004784 \end{array}$$

and $\sqrt{\text{var}(d)} = 3.114888$, so

$$D_T = \frac{14.857143 - 17.959184}{3.114888} = -0.995875$$

The negative value of D_T is caused by an excess of rare alleles. However, from Table 2 on page 592 of Tajima (1989), we see that the 90% confidence interval for D_T in the case $n = 7$ is $(-1.498, 1.728)$, so this value of D_T is not very unusual.

2.4.2 Fu and Li's D

Fu and Li (1993) used the difference $\theta_W - \theta_{FL}$ as a test statistic, or what is essentially the same, $d = S_n - h_n \eta_1$, where $h_n = \sum_{i=1}^{n-1} 1/i$. Again, this statistic is normalized by dividing by the square root of an estimate of $\text{var}(d)$. The ingredients necessary for the computation of $\text{var}(d)$ are given in the previous section, but we skip the somewhat messy details of the derivation, which is similar to the computation of Tajima's denominator.

$$D_{FL} = \frac{S_n - h_n \eta_1}{\sqrt{u_D S_n + v_D S_n^2}} \quad (2.31)$$

where $u_D = h_n - 1 - v_D$, $g_n = \sum_{i=1}^{n-1} 1/i^2$,

$$v_D = 1 + \frac{h_n^2}{g_n + h_n^2} \left(c_n - \frac{n+1}{n-1} \right) \quad \text{and} \quad c_n = \frac{2nh_n - 4(n-1)}{(n-1)(n-2)}$$

To save you some arithmetic, Table 1 of Fu and Li (1993) gives the values of h_n (which they call a_n) and v_D .

Table 2 of Fu and Li (1993) gives cutoff values for their test, which are based on simulation. Since the cutoff values depend on θ , they chose to present conservative percentage points that are valid for $\theta \in [2, 20]$. To illustrate their method, we will consider a data set.

Example 2.2. Hamblin and Aquadro (1996) studied DNA sequence variation at the *glucose dehydrogenase* (*Gld*) locus in *Drosophila simulans*. The *Gld* locus is near the centromere of chromosome 3 in a region of low recombination.

Hamblin and Aquadro sequenced 970 nucleotides from exon 4 from 11 *D. simulans* chromosomes sampled from a population in Raleigh, N.C. These 11 sequences and that of one *D. melanogaster* individual are given in the table below. As usual, dots in rows 1–11 indicate that the sequence agrees with the *D. melanogaster* sequence.

These two *Drosophila* species diverged about 2.5 million years ago, which is about 25 million generations. Since a typical estimate of the *Drosophila* effective population size is one million, it seems likely that the most recent common ancestor of the 11 *D. simulans* individuals, which has mean roughly 4 million generations, will occur before coalescence with the *D. melanogaster* lineage. Thus, the *D. melanogaster* sequence gives us information about the state of the most recent common ancestor of the *D. simulans* individuals and allows us to conclude which nucleotides represent mutations. Note that in all cases but position 5413, the nucleotide in *D. melanogaster* agrees with one of the *D. simulans* individuals. In this case, the state of the most recent common ancestor is ambiguous, but it is clear from the data that the mutation is

```

444444444455555555555555555555
66677789901111233344444555
01925869913579056912568115
92039316947867002235516024
mel CCTTACCCGTGAAGTCCCCTGACCGG
 1 T.GT.G....AG.A....G.....
 2 T.G..G....AG.A....G.....
 3 T....G....G.A....G.A...G.
 4 T....G....AG.A....G.A...G.
 5 T..CT....AGGA...TA....G.
 6 T..CT....G.A...TA....G.
 7 T..CT....AGGA.TTTA....G.
 8 ...CTG.AAC.G.C.TTTGC...A..
 9 ...CTG.AACAGGC...TGC...A..
10 .T...GA..CA...ATTG.AT..GA
11 .T...GA..CA...ATTG.ATT.GA

```

There are 26 segregating sites, but since there are two mutations at 5197, the total number of mutations $\eta = 27$. This represents a deviation from the infinite sites model, but in light of the birthday problem calculations in Section 1.4 is not an unexpected one, since the expected number of nucleotides hit twice is $(27 \cdot 26)/(2 \cdot 970) = 0.35926$ and the probability of no double hit is $\approx \exp(-0.35962) = 0.6982$. However, as we will now see, it is very unusual for there to be only one external mutation at 5486. Table 2 of Fu and Li (1993) gives $h_{11} = 2.929$ and $v_D = 0.214$, so $u_D = h_{11} - 1 - v_D = 1.929 - 0.214 = 1.715$. The value of Fu and Li's statistic is thus

$$D_{FL} = \frac{27 - 1 \cdot 2.929}{\sqrt{(1.715)(27) + (0.214)(27)^2}} = 1.68$$

When $n = 11$, a 95% confidence interval for D_{FL} is $(-2.18, 1.57)$. Thus, there is a significant deficiency of external mutations.

To perform Tajima's test on these data, we ignore the column with two mutations so that there are 25 segregating sites. The folded site frequency spectrum is

$$\begin{array}{rcccccc} m & 1 & 2 & 3 & 4 & 5 \\ \tilde{\eta}_m & 1 & 11 & 4 & 7 & 2 \end{array}$$

Using the fact that $\binom{11}{2} = 55$, we can compute that

$$\theta_\pi = \frac{1 \cdot 10 + 11 \cdot 18 + 4 \cdot 24 + 7 \cdot 28 + 2 \cdot 30}{55} = \frac{560}{55} = 10.181818$$

Dividing θ_π by the 970 bp gives an estimate of π of 0.0105, which is consistent with the estimates for *Adh*, *Amy*, and *rosy* discussed in Section 1.4. In contrast, the estimate of θ based on the 25 segregating sites is $25/2.928968 = 8.535429$. Computing as in the previous example, we find

$$\begin{array}{ll} a_1 = 2.928968 & a_2 = 1.549768 \\ b_1 = 0.400000 & b_2 = 0.272727 \\ c_1 = 0.058583 & c_2 = 0.049884 \\ e_1 = 0.020001 & e_2 = 0.004925 \end{array}$$

and $\sqrt{\widehat{\text{var}}(d)} = 1.858779$, so Tajima's D is

$$D_T = \frac{10.1818181 - 8.535429}{1.858779} = 0.885737$$

Consulting Tajima's table, we see that a 90% confidence interval for D_T is $(-1.572, 1.710)$, so this value of D_T is far from significant.

p values. Tajima's use of the beta distribution and Fu and Li's choice of $\theta \in [2, 20]$ are somewhat arbitrary. To improve the computation of p values for these tests, Simonsen, Churchill, and Aquadro (1995) used (1.25) to construct a $1 - \beta$ confidence interval (θ_L, θ_U) for θ and then defined an interval for the test statistic which had probability $\geq 1 - (\alpha - \beta)$ for a grid of values in $[\theta_L, \theta_U]$.

2.4.3 Fay and Wu's H

Fay and Wu (2000) considered the difference

$$H = \theta_H - \theta_\pi$$

In their original paper, they did not normalize the statistic to have variance approximately 1, and they determined the distribution of H using simulations of a neutral coalescence algorithm (without recombination) conditioning on the observed number of segregating sites. Their motivation for defining H ,

which we will not be able to explain until we consider hitchhiking in Section 6.5, and the site frequency spectrum in Section 7.11, is that the fixation of advantageous alleles should result in an excess of high frequency derived alleles. Since

$$\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \eta_i i^2 \quad \text{versus} \quad \theta_\pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \eta_i i(n-i)$$

and $i^2 > i(n-i)$ when $i > n/2$, this will produce positive values of H . For this reason they consider a one-sided test that rejects the null hypothesis if the observed value is too large.

In computing θ_H an outgroup is needed to infer the ancestral and derived states, but a mutation before the most recent common ancestor or in the outgroup lineage would lead to misinference of the ancestral state. To compensate for this, the probability of misinference was incorporated in the null distribution of the H statistic by exchanging the frequencies of the derived and the ancestral states with probability pd , where d is the net divergence or the average number of fixed differences per site between the two species. If all mutation rates are equal $p = 1/3$. For *Drosophila* data they use $p = 3/8$. This number is chosen based on the idea that in *Drosophila* transitions $A \leftrightarrow G$ and $C \leftrightarrow T$ occur at twice the rate of transversions (the other eight possible mutations). Since for any nucleotide there are two transversions and one transition, $1/2$ of the mutations are transitions and $1/2$ are transversions. Taking into account the rate of back mutations in the two cases, we get

$$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{8}$$

Example 2.3. Accessory gland proteins. This is a group of specialized proteins in the seminal fluid of *Drosophila* that have been suggested to be involved in egg-laying stimulation, remating inhibition, and sperm competition. Among the *Acp* genes, *Acp26Aa* and nearby *Acp26Ab* have been extensively studied. Here we use data of Tsaur, Ting, and Wu (1998) who analyzed 49 sequences from five populations in four continents.

The site frequency spectrum for *Acp26Aa* given in their Table 2 for the 31 sites (out of 38) where the derived nucleotide could be unambiguously inferred with reference to the three outgroup species had $\eta_1 = 9$, $\eta_2 = 5$, $\eta_3 = 2$, $\eta_{46} = 3$, and $\eta_m = 1$ for $m = 6, 7, 11, 16, 21, 29, 31, 38, 39, 42, 45$, and 47. There are 31 segregating sites and $h_{49} = 4.458797$, so

$$\theta_W = 31/4.458797 = 6.95248$$

Using the site frequency spectrum, one can compute

$$\theta_\pi = 5.265306 \quad \theta_H = 15.359694$$

Fay and Wu's test can be run from

<http://www.genetics.wustl.edu/jflab/htest.html>

Because of results of Zeng, Shi, Fu, and Wu (2006) they now scale the statistic by the variance of the difference, so their $H = -4.369243$. Using 10,000 simulations and a back mutation probability of 0.03 produces a p value of 0.0254. If we grind through the details of computing Tajima's D , we find

$$\begin{array}{ll} a_1 = 4.458797 & a_2 = 1.624316 \\ b_1 = 0.347222 & b_2 = 0.231765 \\ c_1 = 0.122946 & c_2 = 0.080038 \\ e_1 = 0.027574 & e_2 = 0.003722 \end{array}$$

and $\sqrt{\text{var}(d)} = 2.077509$, so Tajima's D is

$$D_T = \frac{5.265306 - 6.95248}{2.077509} = -0.812146$$

Consulting Tajima's table, we see that a 90% confidence interval for D_T is $(-1.571, 1.722)$, so this value of D_T is far from significant. Indeed, the output from Fay's program says that the one-tailed p value is 0.229700.

One can, of course, also calculate Fu and Li's D_{FL} . To do this, we begin by computing $v_D = 0.137393$ and $u_D = 3.321404$. Since $\eta_1 = 9$,

$$D_{FL} = \frac{31 - 9 \cdot 4.458797}{\sqrt{(3.321404)(31) + (0.137393)(31)^2}} = -0.595524$$

When $n = 50$, a 95% confidence interval for D_{FL} is $(-1.96, 1.37)$, so again this value is far from significant.

Example 2.4. In the Hamblin and Aquadro (1996) data, throwing out the column in which the outgroup sequence differs from both nucleotides in D . *simulans*, we have 24 segregating sites and a site frequency spectrum of

$$\begin{array}{cccccccccccc} m & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \eta_m & 1 & 10 & 1 & 4 & 2 & 0 & 3 & 2 & 1 & 0 \end{array}$$

Running Fay and Wu's test from the web page with 10,000 simulations and a back mutation probability of 0.03, gives a value for H of 0.117960 (scaled by the variance) and a one-tailed p value of 0.3343.

2.4.4 Conditioning on S_n

Since the number of segregating sites is observable and θ is not, it is natural, as Fay and Wu (2002) have done, to perform simulations of the coalescent conditional on the number of segregating sites $S_n = k$. An easy, but incorrect, way to do this is to generate a genealogy and then place k mutations on it at random. This method was suggested by Hudson (1993, page 27). However,

discussion in his paper indicates he knew it was not correct. The reason is that for fixed values of θ and k not all genealogies have an equal probability of producing k segregating sites.

To see this, let τ_j be the amount of time there are j lineages:

$$P(\tau_2 = t_2, \dots, \tau_n = t_n) = \prod_{j=2}^n \binom{j}{2} \exp\left(-\binom{j}{2} t\right)$$

and note that if $\tau = 2\tau_2 + \dots + n\tau_n$ is the total size of the tree,

$$P(S_n = k | \tau_2 = t_2, \dots, \tau_n = t_n) = e^{-\theta\tau} \frac{(\theta\tau)^k}{k!}$$

Combining the last two formulas and dividing by $P(S_n = k)$, we have

$$\begin{aligned} P(\tau_2 = t_2, \dots, \tau_n = t_n | S_n = k) \\ = c_{\theta, n, k} (2t_2 + \dots + nt_n)^k \prod_{j=2}^n \binom{j}{2} \exp\left(-\left(\binom{j}{2} + j\theta\right) t_j\right) \end{aligned}$$

where $c_{\theta, n, k}$ is a normalizing constant that depends on the indicated parameters. It is clear from the last formula that, for fixed k , as θ changes not only the total size of the tree τ changes but:

- Due to the $j\theta$ in the exponential, the relative sizes of τ_2, \dots, τ_n change, so the distribution of test statistics conditional on $S_n = k$ will depend on θ .
- Due to the $(2t_2 + \dots + nt_n)^k$, the τ_j are no longer independent, since the joint density function is not a product $f_2(t_2) \dots f_n(t_n)$.

Markovstova, Marjoram, and Tavaré (2001) have shown for two tests of neutrality of Depaulis and Veuille (1998) that under the simple but incorrect algorithm of generating a tree and then putting a fixed number of mutations on it, the fraction of observations that fall in an interval claimed to have probability 0.95 can be very small for extreme values of θ . In the other direction, Wall and Hudson (2001) have shown that these problems do not occur if the true value of θ is near Watterson's estimator $\theta_W = S_n/h_n$.

2.5 The HKA test

Suppose now that we have a sample from one species and one sequence from another closely related species. The ratio of the number of segregating sites in one species to the amount of divergence between the two species is determined by the time since divergence of the two species, the effective population size, and the size of the sample, but does not depend on the mutation rate at the locus. Hence, these ratios should be similar for different loci, and sufficiently large differences provide evidence for nonneutral evolution.

Having explained the motivation behind the HKA test, we turn now to the mechanics. Consider data collected from two species, referred to as species A and species B , and from $L \geq 2$ regions of the genome referred to as locus 1 through locus L . Assume that a random sample of n_A gametes from species A have been sequenced at all L loci and n_B gametes from species B have been sequenced at the same loci. Let S_i^A denote the number of sites that are polymorphic at locus i in the sample from species A . Similarly, let S_i^B denote the number of sites that are polymorphic at locus i in the sample from species B . Let D_i denote the number of differences between a random gamete from species A and a random gamete from species B . The $3L$ observations S_i^A , S_i^B , and D_i constitute the data with which the test devised by Hudson, Kreitman, and Aguadé (1987) is carried out.

It is assumed that each locus evolves according to the standard Wright-Fisher infinite sites model: (1) generations are discrete, (2) all mutations are selectively neutral, (3) the number of sites at each locus is very large, so that each mutation occurs at a previously unmutated site, (4) in each generation, mutations occur independently in each gamete and at each locus, (5) at locus i , the number of mutations per gamete in each generation is Poisson distributed with mean u_i , and (6) no recombination occurs within the loci. In addition, we assume that (7) all loci are unlinked, (8) species A and B are at stationarity at the time of sampling with population sizes $2N$ and $2Nf$, respectively, and (9) the two species were derived T' generations ago from a single ancestral population with size $2N(1+f)/2$ gametes.

Letting $\theta_i = 4Nu_i$ and $C(n) = \sum_{j=1}^{n-1} 1/j$, it follows from (1.20) that

$$E(S_i^A) = \theta_i C(n_A) \quad E(S_i^B) = f\theta_i C(n_B)$$

Using (1.22) in Chapter 1 and letting $C_2(n) = \sum_{j=1}^{n-1} 1/j^2$, we have

$$\begin{aligned} \text{var}(S_i^A) &= E(S_i^A) + \theta_i^2 C_2(n_A) \\ \text{var}(S_i^B) &= E(S_i^B) + (f\theta_i)^2 C_2(n_B) \end{aligned}$$

To compute the expected value of D_i , we note that it is $2u_i$ times the expected coalescence time of two individuals: one chosen at random from A and one from B . Those two lineages must stay apart for T' units of time and then coalescence occurs as in a single population of size $2N(1+f)/2$. Measured in units of $2N$ generations, the second phase takes an exponentially distributed amount of time with mean $(1+f)/2$, so letting $T = T'/2N$,

$$ED_i = \theta_i(T + (1+f)/2)$$

To compute the variance, we note that in the first phase, the number of mutations is Poisson with mean $2u_i T' = \theta_i T$. By (1.22) with $n = 2$, the number in the second phase has variance $\theta_i(1+f)/2 + (\theta_i(1+f)/2)^2$ and is independent of the number in the first phase, so

$$\text{var}(D_i) = ED_i + (\theta_i(1 + f)/2)^2$$

There are $L + 2$ parameters. These can be estimated by solving the following $L + 2$ equations:

$$\begin{aligned} \sum_{i=1}^L S_i^A &= C(n_A) \sum_{i=1}^L \hat{\theta}_i \\ \sum_{i=1}^L S_i^B &= C(n_B) \hat{f} \sum_{i=1}^L \hat{\theta}_i \\ \sum_{i=1}^L D_i &= (\hat{T} + (1 + \hat{f})/2) \sum_{i=1}^L \hat{\theta}_i \end{aligned}$$

and for $1 \leq i \leq L - 1$

$$S_i^A + S_i^B + D_i = \hat{\theta}_i \left\{ \hat{T} + (1 + \hat{f})/2 + C(n_A) + C(n_B) \right\}$$

These equations may look complicated, but they are simple to solve. The first can be used to compute $\sum_{i=1}^L \hat{\theta}_i$, the second can then be used to find \hat{f} , the third to compute \hat{T} , and then the individual $\hat{\theta}_i$ can be computed from the remaining $L - 1$. We do not need the equation with $i = L$ since we have already computed the sum of the $\hat{\theta}_i$.

To measure the goodness of fit of these parameters, we can use

$$\begin{aligned} X^2 &= \sum_{i=1}^L (S_i^A - \hat{E}(S_i^A))^2 / \widehat{\text{var}}(S_i^A) \\ &\quad + \sum_{i=1}^L (S_i^B - \hat{E}(S_i^B))^2 / \widehat{\text{var}}(S_i^B) \\ &\quad + \sum_{i=1}^L (D_i - \hat{E}(D_i))^2 / \widehat{\text{var}}(D_i) \end{aligned}$$

If the quantities S_i^A , S_i^B , and D_i were stochastically independent of each other and normally distributed, then the statistic X^2 should be approximately χ^2 with $3L - (L + 2) = 2L - 2$ degrees of freedom. For n_A , n_B , and T sufficiently large, all of these quantities are approximately normally distributed. Since the loci are unlinked, S_i^A is independent of S_j^A and S_j^B when $j \neq i$. Also, S_i^A is independent of S_i^B as long as T is large enough so that there are no shared polymorphisms. However, a small positive correlation is expected between S_i^A and D_i , and between S_i^B and D_i , because a positive fraction of the mutations that contribute to polymorphism also contribute to differences between species. The last observation, and the fact that the normality is only asymptotic, forces the test to be carried out by doing simulations with the estimated parameters.

Example 2.5. Adh. The first application of the HKA test was to the alcohol dehydrogenase locus in *Drosophila melanogaster*. The polymorphism data came from a four-cutter restriction enzyme survey of 81 isochromosomal lines of *D. melanogaster* studied by Kreitman and Aguadé (1986a,b). Nine polymorphic restriction sites were identified in the flanking region and eight in the *Adh* locus. They estimated the effective number of sites to be 414 in the flanking region and 79 in the *Adh* locus. Their interspecific data was based on a sequence comparison of one *D. melanogaster* sequence and one *D. sechellia* sequence. This comparison revealed 210 differences in 4052 bp of flanking sequence and 18 differences in 324 bp in the *Adh* locus. The next table summarizes the data:

	within <i>D. melanogaster</i>			between species		
	sites	variable	%	sites	variable	%
flanking region	414	9	0.022	4052	210	0.052
<i>Adh</i> locus	79	8	0.101	324	18	0.056

Note that the divergence between species is almost the same in the two regions, but there is a considerably higher rate of polymorphism in the *Adh* locus compared to the flanking sequence.

We have no data on the variability within *D. simulans*, so we will suppose that the ancestral population size is the same as the current population size, that is, $f = 1$. To take account of the differing number of sites in the comparisons within ($w_1 = 414$, $w_2 = 79$) and between ($b_1 = 4052$, $b_2 = 324$) species and to prepare for the fact that in the next example the two sample sizes will be different (here $n_1 = n_2 = 81$), we let μ_i be the per nucleotide mutation rate at the i th locus, let $\pi_i = 4N\mu_i$, and note that

$$\begin{aligned}
 ES_1^A &= C(n_1) \cdot w_1 \pi_1 \\
 ES_2^A &= C(n_2) \cdot w_2 \pi_2 \\
 ED_1 &= b_1 \pi_1 (T + 1) \\
 ED_2 &= b_2 \pi_2 (T + 1)
 \end{aligned}$$

Adding the equations as before, we arrive at

$$\begin{aligned}
 S_1^A + S_2^A &= C(n_1) \cdot w_1 \hat{\pi}_1 + C(n_2) \cdot w_2 \hat{\pi}_2 \\
 D_1 + D_2 &= (b_1 \hat{\pi}_1 + b_2 \hat{\pi}_2)(\hat{T} + 1) \\
 S_1^A + D_1 &= C(n_1) \cdot w_1 \hat{\pi}_1 + b_1 \hat{\pi}_1 (\hat{T} + 1)
 \end{aligned}$$

These equations are not as easy to solve as the previous ones. Letting $x = \hat{\pi}_1$, $y = \hat{\pi}_2$, and $z = \hat{T} + 1$, they have the form

$$\begin{aligned}
 c &= ax + by \\
 f &= dxz + eyz \\
 i &= gx + h x z
 \end{aligned}$$

The three equations can be written as

$$z = \frac{f}{dx + ey} \quad y = \frac{c - ax}{b} \quad (1 - gx)\frac{f}{z} = fhx$$

Using the first two in the third equation leads to $\alpha x^2 + \beta x + \gamma = 0$, where

$$\begin{aligned} \alpha &= g \left(d - \frac{ea}{b} \right) \\ \beta &= \frac{gec}{b} + hf - i \left(d - \frac{ea}{b} \right) \\ \gamma &= \frac{-iec}{b} \end{aligned}$$

At this point, there are two cases to consider. If $n_1 = n_2$, $b_1 = w_1$, and $b_2 = w_2$, then

$$d - \frac{ea}{b} = b_1 - \frac{b_2 C(n_1) w_1}{C(n_2) w_2} = 0$$

In this case, $\alpha = 0$ so we solve the linear equation to get $x = -\gamma/\beta$. When $\alpha \neq 0$, the root of the quadratic equation that we want is

$$x = \frac{-\beta + \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}$$

In either case, once x is found, we can compute y and z .

Carrying out the arithmetic in this example gives

$$\hat{\pi}_1 = 6.558 \times 10^{-3}, \quad \hat{\pi}_2 = 8.971 \times 10^{-3}, \quad \hat{T} = 6.734$$

Using the relationships

$$\begin{aligned} \text{var}(S_i^A) &= ES_i^A + (w_i \pi_i)^2 C_2(n_i) \\ \text{var}(D_i) &= ED_i + (b_i \pi_i)^2 \end{aligned}$$

we can compute $X^2 = 6.09$. Monte Carlo simulations with the parameters set equal to these estimates show that the probability of $X^2 > 6.09$ is approximately 0.016. As the reader may have noticed, the flanking sequence is not far enough from the gene region to make it reasonable to assume that the two are unlinked. However, the positive correlation that results from interlocus linkage will shift the distribution of X^2 toward smaller values and make rejections based on the model conservative. Likewise, the intralocus recombination we are ignoring will reduce the variance of the quantities estimated and tend to decrease the value of X^2 .

Having identified a significant departure from neutrality, the next step is to seek an explanation. The fact that there is more polymorphism in the coding region than in the adjacent flanking sequence suggests that something is acting there to make the genealogies larger than they would be under the neutral model. In Section 6.2, we will see that one possible explanation for this is balancing selection acting on the fast/slow polymorphism.

Example 2.6. Drosophila fourth chromosome. Berry, Ajioka, and Kreitman (1991) studied a 1.1kb fragment of the *cubitus interruptus Dominant* (ci^D) locus on the small nonrecombining fourth chromosome for 10 lines of *Drosophila melanogaster* and 9 of *Drosophila simulans*. They found no polymorphism within *Drosophila melanogaster* and a single polymorphism within *Drosophila simulans*. To perform the HKA test, they used data on the 5' region of *Adh* from 11 sequences of Kreitman and Hudson (1991) as their comparison neutral locus. This yielded the following data:

	ci^D	5' <i>Adh</i>
nucleotides	1106	3326
polymorphism	0	30
divergence	54	78

Calculating as in the previous example we find

$$\hat{\pi}_1 = 3.136 \times 10^{-3} \quad \hat{\pi}_2 = 2.072 \times 10^{-2} \quad \hat{T} = 11.74$$

and $X^2 = 6.85$. Using the result of Hudson, Kreitman, and Aguadé (1987) that in this case the statistic has approximately a chi-square distribution with 1 degree of freedom, Berry, Ajioka, and Kreitman (1991) concluded that the probability of an X^2 value this large is < 0.01 . (Note that the value of 1 here contrasts with the $2L - 2 = 2$ degrees of freedom that the statistic would have if S_i^A and D_i were independent.)

One explanation for these data is purifying selection. The original population sizes in both species were small, permitting effectively neutral drift of mildly deleterious alleles and causing the accumulation of fixed differences between the two species. Subsequent population expansion has increased the efficacy of selection against such mildly deleterious mutations, and what we see, within species, is the wholesale removal of variation by purifying selection. While this explanation is possible, it seems unlikely. Given the lack of variation at both silent and replacement sites, a slightly deleterious allele model would require that selection coefficients against both silent and replacement sites would fall between $1/2N_2$ and $1/2N_1$, where N_1 and N_2 are the pre- and post-expansion population sizes. It is unlikely that these two types of mutations, which have entirely different functional consequences, would have similar selection coefficients.

A second explanation is that a selective sweep eliminated variation in this region for both species. In order to estimate the time of occurrence of such a sweep, we note that if T_{tot} is the total time in the genealogy of our sample, μ is the mutation rate per nucleotide per generation, and k is the number of silent sites, then the expected number of segregating sites

$$ES = T_{tot}\mu k$$

To simplify calculations, we will suppose that the sweep was recent enough so that the resulting genealogy is star-shaped. In this case, $T_{tot} = nt$, where n is

the sample size and t is the time of the sweep. For the *Drosophila melanogaster* sample, $S = 0$, so we are left with an estimate of $t = 0$. For *D. simulans* substituting 1 for ES , and taking $n = 9$, $k = 331$, and $\mu = 1 \times 10^{-9}$, we arrive at

$$t = \frac{ES}{nk\mu} = \frac{1}{9 \cdot 331 \cdot 10^{-9}} = 3.35 \times 10^5 \text{ generations ago}$$

Assuming 10 generations per year, this translates into 33,500 years.

Having assumed a star-shaped phylogeny and calculated a time, we should go back and check to see if our assumption is justified. The probability of no coalescence in a sample of size n during t generations in a population of size N is

$$\approx \exp\left(-\binom{n}{2} \frac{t}{2N}\right)$$

If we take $2N = 5 \times 10^6$, $n = 9$, and $t = 3.35 \times 10^5$, then the above

$$= \exp\left(-36 \frac{3.35}{50}\right) = e^{-2.412} = 0.0896$$

i.e., it is very likely that there has been at least one coalescence. Once one abandons the assumption of a star-shaped phylogeny, calculations become difficult and it is natural to turn to simulation. Using $4N\mu = 3$ for *D. simulans*, Berry, Ajioka, and Kreitman (1991) computed that there was a 50% probability of sweep in the last $0.36N$ generations, or 72,000 years.

2.6 McDonald-Kreitman test

To describe the test of McDonald and Kreitman (1991), we need some notation. Of M possible mutations in a coding region, let M_r be the number of possible neutral replacement mutations (i.e., ones that change the amino acid but not the effectiveness of the protein) and let M_s be the number of possible neutral synonymous mutations. By definition, all of the $M - M_r - M_s$ remaining mutations are deleterious.

Let μ be the mutation rate per nucleotide, so that the mutation rate for any one of the three possible changes at a site is $\mu/3$. Under the neutral theory, the expected number of fixed replacement substitutions in a set of alleles is $T_b(\mu/3)M_r$, where T_b is the total time on between-species branches. The expected number of fixed synonymous substitutions in a set of alleles is $T_b(\mu/3)M_s$. For a particular phylogeny and mutation rate, the number of replacement substitutions is independent of the number of synonymous substitutions. Therefore, the ratio of expected replacement to expected synonymous fixed mutations is

$$\frac{T_b(\mu/3)M_r}{T_b(\mu/3)M_s} = \frac{M_r}{M_s}$$

If T_w is the total time on within-species branches, then the ratio of expected replacement to expected synonymous polymorphic mutations is

$$\frac{T_w(\mu/3)M_r}{T_w(\mu/3)M_s} = \frac{M_r}{M_s}$$

Thus, if protein evolution occurs by neutral processes, the two ratios are the same and we can use standard statistical tests for 2×2 contingency tables to test this null hypothesis.

Under the alternative model of adaptive protein evolution, there should be relatively more replacement substitution between species than replacement polymorphism within a species, so a deviation in this direction is interpreted as evidence for positive selection.

Example 2.7. To explain the workings of the test, we will begin with the original data set of McDonald and Kreitman (1991). They compared DNA sequences of the *Adh* locus in *Drosophila melanogaster*, *D. simulans*, and *D. yakuba*. The DNA sequence data can be found on page 653 of their paper. To carry out the test, the following summary is sufficient:

	Fixed Polymorphic	
Replacement	7	2
Synonymous	17	42

To analyze the table, we first compute the number of observations we expect to find in each cell (given in parentheses in the next table):

	Fixed	Polymorphic	Total
Replacement	7 (3.176)	2 (5.824)	9
Synonymous	17 (20.824)	42 (38.176)	59
Total	24	44	68

Then we compute the χ^2 statistic:

$$\frac{(7 - 3.176)^2}{3.176} + \frac{(2 - 5.824)^2}{5.824} + \frac{(17 - 20.824)^2}{20.824} + \frac{(42 - 38.176)^2}{38.176} = 8.198$$

The number of degrees of freedom in this case is 1, so the χ^2 distribution is just the square of a standard normal, χ , and we can use a table of the standard normal to conclude that the probability of a deviation this large by chance is $2P(\chi > \sqrt{8.198}) = 0.0042$. McDonald and Kreitman analyzed the contingency table with a G test of independence (with the Williams correction for continuity), finding $G = 7.43$ and $p = 0.006$.

Geneticists have embraced the McDonald-Kreitman test as a useful tool for looking for positive selection. However, the initial paper did not get such a warm reception. Graur and Li, and Whittam and Nei, each wrote letters that appeared in the November 14, 1991 issue of *Nature* suggesting that the test had serious problems. Both pairs of authors objected to some of the bookkeeping involved in the three-species comparison. For this reason, we will now consider only pairwise comparisons. The data for *D. melanogaster* and *D. simulans*, is

[illegible]

The contingency table is now much different, with the 24 fixed differences having been reduced to just 4.

	Fixed	Polymorphic	Total
Replacement	2	2	4
Synonymous	2	26	28
Total	4	28	32

Since the cell counts are small, we analyze the results with *Fisher's exact test*. To derive this test, we note that if we condition on the number of replacement substitutions n_r , then the number of fixed replacements, n_{fr} , is binomial(n_r, p), where $p = T_b / (T_w + T_b)$. Likewise, if we condition on the number of synonymous substitutions, n_s , then the number of fixed synonymous substitutions, n_{fs} , is binomial(n_s, p). Let n_f and n_p be the number of fixed and polymorphic substitutions. The probability of a given table conditioned on the marginal values n_r, n_s, n_f, n_p is

$$\frac{n_r!}{n_{fr}!n_{pr}!}p^{n_{fr}}(1-p)^{n_{pr}} \cdot \frac{n_s!}{n_{fs}!n_{ps}!}p^{n_{fs}}(1-p)^{n_{ps}} = \frac{C}{n_{fr}!n_{pr}!n_{fs}!n_{ps}!}$$

where C is a constant independent of $(n_{fr}, n_{pr}, n_{fs}, n_{ps})$.

There are only five 2×2 tables with the indicated row and column sums: n_{fr} can be 0, 1, 2, 3, or 4 and this determines the rest of the entries. Of these, the ones with $n_{fr} = 2, 3, 4$ are more extreme than the indicated table. Using the preceding formula, it is easy to compute the conditional probability that $n_{fr} = k$ given the row and column sums:

k	0	1	2	3	4
prob.	0.569383	0.364405	0.063070	0.003115	0.000028

From this we see that the probability of a table this extreme is $0.066212 > 0.05$.

In contrast, if we compare *D. simulans* with *D. yakuba* using the data in McDonald and Kreitman (1991),

```
111111111111111111111111111111111111
788899001222233444455555555555555555
8137576890237801234590125556991
1640048993951346513204871570064
GGTCGGCCCCTCACCCCTTCACCCCGCGC
.TC..T.A.....T..
.TC.A..A...T.....T.G
.T...T.A.T.....T..
.T...T.A...T..T.....T..A...
.T...T.A.....T.....
.TC..T.A.....
RSSSSSRSSSSSSSSSSSRSSSSSRSSSS
PPFPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
```

The contingency table is:

	Fixed	Polymorphic	Total
Replacement	6	0	6
Synonymous	17	29	46
Total	23	29	52

Fisher’s exact test gives that the probability $n_{fr} = 6$ given the row and column sums is 0.00493, so there is a clear departure from neutral evolution.

Example 2.8. Eanes, Kirchner, and Yoon (1993) sequenced 32 and 12 copies of the gene (*G6pd*) in *Drosophila melanogaster* and *D. simulans* respectively. This revealed the following results (the number of observations we expect to find in each cell is given in parentheses):

	Fixed	Polymorphic	Total
Replacement	21 (12.718)	2 (10.282)	23
Synonymous	26 (34.282)	36 (27.717)	62
Total	47	38	85

The χ^2 statistic is 16.541. The probability of a χ^2 value this large by chance is < 0.0001 . Thus, there is a very strong signal of departure from neutral evolution. The most likely explanation is that replacement substitutions are not neutral but have been periodically selected through the populations of one or both species as advantageous amino acid mutations.

Example 2.9. Accessory gland proteins are specialized proteins in the seminal fluid of *Drosophila*. They have been suggested to be involved in egg-laying stimulation, remating inhibition, and sperm competition, so there is reason to suspect that they are under positive selection. Tsaur, Ting, and Wu (1998) studied the evolution of *Acp26Aa*. They sequenced 39 *D. melanogaster* chromosomes, which they combined with 10 published *D. melanogaster* sequences

and 1 *D. simulans* sequence in Aguadé, Miyashita, and Langley (1992). The reader's first reaction to the sample size of 1 for *D. simulans* may be that this makes it impossible to determine whether sites are polymorphic in *D. simulans*. This does not ruin the test, however. It just reduces T_w to the total time in the genealogy for the *D. melanogaster* sample.

The next table gives the data as well as the number of observations we expect to find in each cell (given in parentheses):

	Fixed	Polymorphic	Total
Replacement	75 (69.493)	22 (27.507)	97
Synonymous	21 (26.507)	16 (10.492)	37
Total	96	38	134

The χ^2 statistic is 5.574. The probability of a χ^2 value this large by chance is $2P(\chi \geq \sqrt{5.574}) = 0.0181$. It is interesting to note that while the McDonald-Kreitman test leads to a rejection of the neutral model, Tajima's D , which is -0.875 , and Fu and Li's D , which is -0.118 , do not come close to rejection.

Example 2.10. The first three examples have all shown a larger ratio of replacement to silent changes between species. Mitochondrial DNA shows the opposite pattern. Nachman (1998) describes the results of 25 comparisons involving a wide variety of organisms. Seventeen of the contingency tables deviate from the neutral expectation, and most of the deviations (15 of 17) are in the direction of greater ratio of replacement to silent variation within species. A typical example is the comparison of the ATPase gene 6 from *Drosophila melanogaster* and *D. simulans* from Kaneko, Satta, Matsura, and Chigusa (1993). As before, the number of observations we expect to find in each cell is given in parentheses:

	Fixed	Polymorphic	Total
Replacement	4 (1.482)	4 (6.518)	8
Synonymous	1 (3.518)	18 (15.482)	19
Total	5	22	27

The χ^2 statistic is 7.467. The probability of a χ^2 value this large by chance is $2P(\chi \geq \sqrt{7.467}) = 0.0064$. One explanation for a larger ratio of replacement to silent changes within populations is that many of the replacement polymorphisms are mildly deleterious.



<http://www.springer.com/978-0-387-78168-6>

Probability Models for DNA Sequence Evolution

Durrett, R.

2008, XII, 431 p., Hardcover

ISBN: 978-0-387-78168-6