

Prepared by Professor Robert Berwick.

Notes on population genetics and evolution: “Cheat sheet” for review

1. Genetic drift

Terminology. *Genetic drift* is the stochastic fluctuation in allele frequency due to random sampling in a population.

Polymorphism describes sites (nucleotide positions, etc.) variable *within* a species; *divergence* describes sites variable between species.

1.1 Wright-Fisher model.

The Wright-Fisher model describes the process of genetic drift within a finite population.

The model assumes:

1. N diploid organisms (so, $2N$ gametes)
2. Monoecious reproduction with an infinite # of gametes (no sexual recombination)
3. Non-overlapping generations
4. Random mating
5. No mutation
6. No selection

The Wright-Fisher model assumes that the ancestors of the present generation are obtained by random sampling with replacement from the previous generation. Looking forward in time, consider the familiar starting point of classical population genetics: two alleles, A and a , segregating in the population. Let i be the number of copies of allele A , so that $N-i$ is the number of copies of allele a . Thus the current frequency of A in the population is $p = i/N$, and the current frequency of a is $1-p$. We assume that there is no difference in fitness between the two alleles, that the population is not subdivided, and that mutations do not occur. This gives the familiar formula for the probability that a gene with i copies in the present generation is found in j copies in the next generation:

$$P_{ij} = \binom{N}{j} p^j (1-p)^{N-j} \quad 0 \leq j \leq N$$

Let the current generation be generation zero and K_t represent the counts of allele A in future generations. The binomial equation above states that K_1 is binomially distributed with parameters N and $p = i/N$, given $K_0 = i$.

From standard results in statistics, we know the mean and variance of K_1 :

$$E[K_1] = Np = i$$

$$\text{Var}[K_1] = Np(p-1)$$

So, the number of copies of A is expected to remain the same on average, but in fact may take any value from zero to N . A particular variant may become extinct (go to zero copies) or fix (go to N copies) in the population even in a single generation. Over time, the frequency of A will drift randomly according to the Markov chain with transition

probabilities given by the above formula, and eventually one or the other allele will be lost from the population.

Perhaps the easiest way to see how the Wright-Fisher binomial sampling model works is through a biologically motivated example. Imagine that before dying each individual in the population produces a very large number of gametes. However, the population size is tightly controlled so that only N of these can be admitted into the next generation. The frequency of allele A in the gamete pool will be i/N , and because there are no fitness differences, the next generation is obtained by randomly choosing N alleles. The connection to the binomial distribution is clear: we perform N trials, each with $p = i/N$ chance of success. Because the gamete pool is so large, we assume it is not depleted by this sampling, so the probability i/N is still the same for each trial. The distribution of the number of A alleles in the next generation is the binomial distribution with parameters $(N, i/N)$ as expected.

The decay of heterozygosity.

Before we take up the backward, ancestral process for the Wright-Fisher model, we will look at the classical forward derivation. The heterozygosity of a population is defined to be the probability that two randomly sampled gene copies are different. For a randomly mating diploid population, this is equivalent to the chance that an individual is heterozygous at a locus. Let the current generation be generation zero, and let p_0 be the frequency of A now. The heterozygosity of the population now is equal to $H_0 = 2p_0(1-p_0)$, the binomial chance that one allele A (and one a) is chosen in two random draws.

Let the random variable P_t represent the frequencies of A in each future generation t . Then, as we have seen in earlier lectures, in the next generation the heterozygosity will have changed to be $H_1 = 2P_1(1-P_1)$. However, H_1 will vary depending on the random realization of the process of genetic drift. On average, heterozygosity (variation) will be lost through drift:

$$\begin{aligned} E[H_1] &= E[2P_1(1-P_1)] \\ &= 2(E[P_1] - E[P_1]^2 - \text{Var}[P_1]) \\ &= 2p_0(1-p_0)\left(1 - \frac{1}{2N}\right) \\ &= H_0\left(1 - \frac{1}{2N}\right) \end{aligned}$$

In the haploid case, we replace $2N$ by N . After t generations, we have:

$$E[H_t] = H_0 \left(1 - \frac{1}{2N}\right)^t$$

The approximation is valid for large N . Thus, as we've seen, in the Wright-Fisher model, heterozygosity decays at rate $1/N$ per generation, $1/2N$ if diploid. The decrease of

heterozygosity is a common measure of genetic drift, and we say that the drift occurs in the Wright-Fisher model at rate $1/N$ ($1/(2N)$ if diploid) per generation.

We can also get the same result in this way. From the Hardy-Weinberg principle, if p is the frequency of the allele A_1 and $(1-p)$ is the frequency of allele A_2 , and if there is random mating, the frequency of A_1A_1 , A_1A_2 , and A_2A_2 individuals in the next generation is given by p^2 , $2p(1-p)$, and $(1-p)^2$. Thus, the proportion of *homozygous* individuals, $F = p^2 + (1-p)^2$ and the fraction of heterozygous individuals, $H = 2p(1-p)$.

The probability of an individual being homozygous (i.e., 'the same', either A_1A_1 or A_2A_2) in the next generation is as follows (following the recitation/lecture analysis):

$$F_{t+1} = \frac{1}{2N} + (1 - \frac{1}{2N})F_t$$

The probability of an individual being heterozygous, or 'different' is $H = 1 - F$. So,

$$H_{t+1} = [1 - 1/(2N)]H_t$$

$$H_t = [1 - 1/(2N)]^t H_0$$

When $x \ll 1$, then $(1-x)^t$ is approximately e^{-xt} , so,

$$H_t \approx H_0 e^{-t/2N}$$

As $t \rightarrow \infty$, $H_t \rightarrow 0$

A familiar example of genetic drift. Say there ~15,000 genes in the human genome, and since we are diploid, that means we have two copies of each. On average, there is a polymorphic nucleotide site about every 500 to 2000 bases. Let's say 1 kb for argument's sake. Let's further say the average gene is 1000 bases (1 kb). So by this crude reasoning, we are all heterozygous at every locus, on average. Now, suppose you are an only child. You got one copy of a gene from mom and the other copy from dad. Since you are their only child, that means that one allele from each gene in each parent's genome did not make it into the next generation (i.e., was "lost"). Clearly, natural selection did not favor anything like all 15,000 (times two) of the alleles that made it into your genome and disfavor the 15,000 (times two) that did not. Almost all of the alleles that made it into your genome made it at random.

Fluctuations in population size. Suppose the population size is N_1, N_2, \dots, N_t in generations 1, 2, ..., t . Then we have:

$$H_1 = [1 - 1/(2N_1)]H_0$$

$$H_2 = [1 - 1/(2N_2)]H_1 = (1 - 1/(2N_2))(1 - 1/(2N_1))H_0$$

$$H_t = [1 - 1/(2N_t)]$$

$$H_{t-1} = [1 - 1/(2N_t)] \dots [1 - 1/(2N_2)][1 - 1/(2N_1)]H_0$$

Let N_e be the *effective size* of the population, i.e., the size of a population that has the same rate of loss of heterozygosity as the one with fluctuating population sizes. Thus, we want to find the value of N_e that satisfies:

$$[1-1/(2N_t)]H_0=[1-1/(2N_t)]\dots[1-1/(2N_2)](1-1/(2N_1))H_0$$

Again approximating by a Taylor series, we have:

$$e^{-1/2N_e} = (e^{-1/2N_t})\dots(e^{-1/2N_1}) \text{ or, taking logs of both sides:}$$

$$\frac{1}{N_e} = \left(\frac{1}{t}\right) \left(\frac{1}{N_t} + \dots + \frac{1}{N_2} + \frac{1}{N_1}\right)$$

Thus N_e is the harmonic mean of the actual population size. Since the harmonic mean is dominated by the smallest terms, population bottlenecks, or brief reductions in actual population size, can have a strong influence on the effective population size and heterozygosity (read: variance). (This is called *variance effective population size*.)

Effective population size is *crucial* to all the calculations because the mathematical results depend on the assumption that the Wright-Fisher idealization of binomial draws is being maintained.

1.2 Effective population size

More generally, the *effective population size* is thus the size of an *idealized* population that has the same magnitude of drift as an idealized Wright-Fisher population. The effective population size is always less than the census population size due to factors such as this one. Other cases may be dealt with as in the case of fluctuating population size, by figuring out what value of N that would yield the *same rate of loss in heterozygosity* ‘as if’ the population were an ideal Fisher-Wright sample – that is, by calculating, in each particular instance, what the reduction in variance is from generation to generation, as we did above, and then back calculating what the value of N ‘should have been.’ Some examples include:

1. Unequal numbers of males and females. Imagine a zoo population with 20 males and 20 females. Due to the dominance hierarchy only one of the males actually breeds. What is the relevant population size that informs us about the strength of drift in this system? 40? 21? If N_m is the number of breeding males (1 in this example) and N_f the number of breeding females (20), then half of the genes in the offspring generation will derive from parent females and half from parent males
2. Overlapping generations
3. Non-Poisson distribution of fecundity (i.e., different numbers of offspring)
4. Non-random mating, i.e., population structure in general

3. Coalescent theory

Coalescent theory describes the genealogical relationships among individuals in a Wright-Fisher population.

Notation: Let T_2 be the time in generations until the most recent common ancestor (coalescence) of *two* genes (alleles, sequences,...) chosen at random from a population of size N (better and more correct: N_e). We will also call these *lineages*. Also, from now on, whenever we use “ N ” we really mean N_e). We assume in what follows that the genes,

sequences, etc. are drawn from a single species. (This is important for some of the statistical calculations testing for selection, below.)

In general, T_i = the time until the coalescence of i lineages (genes, alleles, sequences,...). That is, after coalescence, the two genes are *identical*. We are interested in the distribution of the ‘waiting times’ until each coalescence, as well as the variance of these times, and, further, the expected waiting time and the total waiting time until *all* lineages have collapsed into a single common ancestor. It turns out that all this can be described as a stochastic process with rather simple properties. Note that each coalescent event is independent of all others – the waiting times are independent.

3.1 Basic results.

Measured in discrete time, in a Wright-Fisher population of size $2N$ the distribution of waiting times until the collapse (coalescence, identity) of two sequences is geometric with the probability of success p (= *coalescence*) = $1/(2N)$ in any one generation, and so the probability of failure (= *not* coalescing) is $1-p$, or $[1-1/(2N)]$. (Note the close relation between this and the heterozygosity computation.) It is easy to see that the waiting times form a geometric distribution by considering the probability that up until time t a coalescent event *has not* occurred, as the product of t ‘not coalescing’ events, just as with the heterozygosity iteration. If we let $P(T_2 > t)$ denote the probability that two lineages have *not* coalesced, for times $t=0, 1, 2, \dots$, then this is simply:

$$P(T_2 > t) = \left(1 - \frac{1}{2N}\right)^t, t = 0, 1, 2, \dots$$

so the probability that two lineages collapse at exactly the t^{th} time step is:

$$\begin{aligned} P(T_2 = t + 1) &= P(T_2 > t) - P(T_2 > t + 1), t = 0, 1, 2, \dots \\ &= \left(1 - \frac{1}{2N}\right)^t - \left(1 - \frac{1}{2N}\right)^{t+1} \\ &= \left(1 - \frac{1}{2N}\right)^t \left(1 - 1 + \frac{1}{2N}\right) \\ &= \frac{1}{2N} \left(1 - \frac{1}{2N}\right)^t \end{aligned}$$

And this is clearly a geometric distribution.

3.1.1 A very, very intuitive picture.

We can gain a very intuitive picture of the same process by the following argument. We start by considering the coalescence time in a sample of two genes. Genes X and Y live in the present generation, and their common ancestor A lived t generations ago. Consequently, as we look backward from the present into the past, the two lines of descent remain distinct for t generations, at which time they coalesce into a single line of descent. In a given generation, the lines coalesce if the two genes in that generation are copies of a single parental gene in the generation before. Otherwise, the two lines remain distinct.

What can we say about the length of time, t , that they remain distinct? The problem is a lot like the following. Suppose that we are talking about the life-span of a piece of kitchen glassware. Eventually, someone will drop it and it will break. Suppose that the probability of breakage is h per day and its expected lifespan is T days. To see how h and T are related, consider the two things that can happen on day one: The glass either survives the first day or it breaks. It breaks with probability h , and in this case its lifespan is 1 day. It survives with probability $1-h$. Further, for surviving glasses, the mean lifespan is $1+T$. Why? Because a glass doesn't age; its hazard of breakage is always h regardless of how old it is. Consequently, the expected life remaining to a glass does not depend on how old it is. Our one-day-old glass can expect to live T additional days, so its expected lifespan is $1+T$. Putting these facts together gives an expression for T in terms of itself:

$$T = h + (1-h)(1+T)$$

So, $T=1/h$. (You can also derive the result using calculus.) Returning to gene lineages, if we knew the 'hazard,' h , that the lines of descent will coalesce (or collide) during a generation, then this would tell us immediately the mean number of generations until the two lineages coalesce. But we do know this: If there are G distinct genes in the population, then $h = 1/G$. More generally, the probability that two genes are identical when drawn from a (diploid) population is $1/(2N)$.

3.1.2 Results derive from the geometric distribution of 'waiting times' until lineages coalesce

A geometric probability distribution may be described by $Prob\{x=i\}=q^{i-1}p$, where p is the probability of success on any one trial, and q is the probability of failure. From basic statistical theory, we know that the mean of a geometric distribution function is just the inverse of the probability of success, $p=1/2N$, and its variance is $q/p^2 = (1-p)/p^2$.

So:

(i) The expected value for the time to coalescence for a sample of 2 genes (sequences,...) is just the following, where N measured in units of generations:

$$E[T_2] = 2N = \frac{2N}{\binom{2}{2}}$$

Further, we also know the variance of the geometric is in this case:

$$\text{var}(T_2) = \frac{2N-1}{\frac{1}{4N^2}} = 2N(N-1) \approx 4N^2$$

Note that the variance is quite large.

In general, for n lineages:

(ii) The expected time to coalescence from k to $k-1$ lineages is:

$$E[T_k] = \frac{4N}{k(k-1)} = \frac{2N}{\binom{k}{2}}$$

So for example, if we have 3 sequences, the time to the first coalescence will be, on average:

$$E[T_3] = \frac{4N}{3(3-1)} = \frac{2N}{\binom{3}{2}} = \frac{1}{3} 2N$$

This makes sense, since for the first coalescence, we have a (3 choose 2) or 3 possible ways of collapsing 3 sequences together (1st and 2nd; 1st and 3rd; 2nd and 3rd) – there are more cars in the intersection, so a higher chance that they will ‘collide’, and so a lower waiting time until they do coalesce (specifically, 1/3 of the average time when there are only 2 sequences).

And so on: for four lineages (sequences), we initially have 4-choose-2 options to collapse, which gives an expected time to first collapse of $2N/6 = 1/6 (2N)$, etc.

(iii) The total length of all the branches in the genealogy tree, $E[T_{tot}]$, which is an important value that we’ll use to figure out the expected nucleotide diversity, may be computed as follows:

$$E[T_{tot}] = \sum_{i=2}^n i E[T_i] = \sum_{i=2}^n i \frac{2N}{\binom{i}{2}} = 4N \sum_{i=1}^{n-1} i$$

(iv) The time to coalescence of all n lineages (the so-called “time to most recent common ancestor,” MRCA), and so the total expected depth of the coalescent, can be found as follows. Note that this expected time is ‘about’ $4N$, a bit less with a small factor dependent on the sample size n . Therefore, sampling an $n+1$ st sequence adds only $2/n$ to what may already be a sizeable number. This has implications for the measurement of DNA sequence polymorphism, which we describe below. Further, the equation for MRCA means that in generational units of $2N$, the time to MRCA is always very close to its asymptotic value of 2, even for moderate n . Thus, for all but the smallest samples, there will likely be a large number of coalescent events in the very recent history of the sample.

$$\begin{aligned} E[T_n] &= 2N \sum_{i=2}^{n-1} \frac{2}{i(i-1)} = 2N \cdot 2 \sum_{i=2}^{n-1} \frac{1}{i-1} - \frac{1}{i} \\ &= 2N \cdot 2 \left(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \dots - \frac{1}{n-1} + \frac{1}{n-1} - \frac{1}{n} \right) \\ &= 2N \cdot 2 \left(1 - \frac{1}{n} \right) \end{aligned}$$

(v) Properties of the shape and size of the coalescent tree.

Note that the full coalescent tree is dominated by the most ancient coalescent, of depth on average $2N$. The tree collapses to just two lineages in expected time $2N$, then collapses all the rest of the way, from 2 lineages to 1 in another expected time of $2N$.

(vi) We can pass from the discrete, geometric distribution to its continuous analog as follows, following the Rice book: Since for $2N > 100$, we can expand $e^{-t/2N}$ as a Taylor series approximately equal to $(1-t/2N)$, we can rewrite the geometric distribution as an exponential distribution:

$$P(T_k > t) \approx \left(1 - \binom{k}{2} \frac{1}{2N}\right)^t, \text{ which as } N \rightarrow \infty$$

$$\approx \frac{\binom{k}{2}}{2N} e^{-\binom{k}{2} \frac{t}{2N}}$$

If we rescale time in generational units of $\tau = t/2N$, so that one ‘clock tick’ is set to this value, then we can simplify the basic coalescent results in a much neater form, which will also let us get the variance in a useful form:

$$P(T_k) = e^{-\binom{k}{2}\tau}$$

$$E(T_k) = \binom{k}{2}^{-1}$$

$$\text{var}(T_k) = \binom{k}{2}^{-2}$$

We see that $2N$ (where N is of course actually the effective population size) is the ‘natural unit’ for considering lineage coalescence.

3.2 Adding mutations: The coalescent and the neutral theory

We now add mutations to the genealogical tree to get some actual results and tests. The idea is this: rather than ask, “for a given mutation parameter, what can we say about the ancestry of the sample?” we ask the more relevant question: “given this sample, what can we say about the population?”

The key idea to adding mutations to the coalescent tree is that what we observe in terms of segregating sites are two superimposed, independent stochastic processes: one due to the lineages collapsing (which are $n-1$ independent, exponentially/geometrically distributed waiting times) and the other due to the random, neutral mutations sprinkled on top of this lineage collapse pattern (which for large population sizes may be considered to be Poisson distributed).

The expected number of segregating sites in a sample of size n , S_n , (what we also simply call polymorphisms), will just be the neutral mutation rate u times the expected time in the coalescent, or:

$$E[S_n] = uE[T_{tot}] = 4Nu \sum_{i=1}^{n-1} i$$

We label $4Nu$ as θ , so we can re-write this as:

$$E[S_n] = a_n \theta \text{ where } a_n = \sum_{i=1}^{n-1} i$$

We now define a statistical estimator of θ as follows:

$$\hat{\theta}_w = \frac{S_n}{a_n}, \text{ where } a_n = \sum_{i=1}^{n-1} i, \text{ in other words}$$

$$\hat{\theta}_w = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n-1}}$$

Here, S_n is what we measure, e.g., from SNP data, while theta is estimated. This particular estimator was first given by Watterson (1975), and it is ‘unbiased’ in the sense that its expected value is the true value of the number of segregating sites.

Of course, in order to do statistical estimation, we really need to know something about the variance of our estimator. The variance of the number of segregating sites is really the sum of *two* components, one due to the coalescence tree (conditioned on the depth of the tree), and one due to the variation due to the Poisson mutation process. We give the following sketch to compute this without the full proof, where T_n is the depth of the coalescence tree:

$$\begin{aligned} \text{Var}(S_n) &= E[\text{Var}(S_n | T_n)] + \text{Var}(E[S_n | T_n]) \\ &= E\left[\frac{\theta}{2} T_n\right] + \text{Var}\left(\frac{\theta}{2} T_n\right) \\ &= \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \end{aligned}$$

It is easy to see that the variance of $\hat{\theta}_w = \text{var}(\hat{\theta}_w) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} = \frac{1}{a_n} \theta + \frac{b_n}{a_n^2} \theta^2$,

so this variance approaches 0 as n approaches infinity. This means one can attain any level of precision desired by choosing the sample size sufficiently large (but don’t expect the precision to be much better than half the size of the estimate unless the sample size is absurdly large – we leave that exercise to you to check out). Such estimators are called *consistent*.

An important point: our model of mutation here is traditionally called the *infinite sites* model. Note that in doing this computation about neutral mutations and their ultimate ‘effect’ in showing up as segregating sites, via sprinkling on the coalescent branches, we have made implicit use of an assumption: each mutation is at a *different* site in the sequence, so that each mutation produces a distinct, segregating ‘spot’ on the DNA sequence. Roughly, this is what permits us to equate the number of segregating sites to the simple multiplication of the neutral mutation rate times the expected tree depth. You might want to think through what would happen if we allowed multiple ‘hits’ at the same nucleotide position. If we assume that the mutation rate is, say, $10^{-6} - 10^{-8}$ per base pair per replication, and that sequences are of ‘average’ length (like what?) then this assumption does not seem too bad, so the infinite sites model seems OK for sequences.

3.3 Using the coalescent to test hypotheses about nucleotide diversity: Tajima’s D

Now we can actually construct a test of the neutral hypothesis, based on two estimators of theta. Another way we have of estimating θ is to just calculate the number of mutations separating individuals two at a time, and average over all pairs. This may be thought of as a sample average to estimate a population average, and is a common measure of nucleotide diversity. Denote by

$$S_{ij} = \text{number of mutations separating individuals } i \text{ and } j$$

Under the infinite sites assumption, we can calculate S_{ij} from a sample by calculating the number of segregating sites between sequences i and j . If we average S_{ij} over all pairs (i,j) in a sample of size n this is called the average number of pairwise differences. We denote this by:

$$D_n = \frac{2}{n(n-1)} \sum_{i \leq j} S_{ij}$$

Note that we can think of individuals (i,j) as a sample of size 2, so:

$$E[S_{ij}] = E[S_2] = \theta$$

and so,

$$E[D_n] = \frac{2}{n(n-1)} \sum_{i \leq j} E[S_{ij}] = \theta$$

Thus, D_n is another, unbiased estimator of θ , called $\hat{\theta}_T$. Tajima (1981) was the first to investigate its properties. He noticed that since $E[D] = \hat{\theta}_T = \theta$ and $E[S_n] = \hat{\theta}_W = a_n \theta$, (a_n as above, i.e., $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$) then the expected value of the difference $\hat{\theta}_T - \hat{\theta}_W$ should be zero under the standard neutral model. Significant deviations from zero should cause the null model to be rejected (i.e., there is possibly positive selection). Specifically, Tajima (1989) proposed the test statistic:

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{Var}[\hat{\theta}_T - \hat{\theta}_W]}}$$

The denominator of Tajima's D is an attempt to normalize for the effect of sample size on the critical values. We have to estimate this denominator (hence the 'hat' on Var) from the data by using the formula:

$$\sqrt{\hat{Var}[\hat{\theta}_T - \hat{\theta}_W]} = e_1 S + e_2 S(S-1)$$

where

$$e_1 = \frac{1}{a_n} \left(\frac{n+1}{3(n-1)} - \frac{1}{a_n} \right), \quad e_2 = \frac{1}{a_n^2 + b_n} \left(\frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{na_n} + \frac{b_n}{a_n^2} \right)$$

where $b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$

This looks formidably complicated, but it's really not (though tricky to derive): the coefficients come from the computation of the variance difference between the two estimators just as we derived the variance of S_n above.

To actually use this test, Tajima suggested that the distribution of D might be approximated by a certain form (not quite a normal distribution, but a beta distribution), and provided tables of critical values for the rejection of the standard neutral model. The upper (lower) critical value is the value above (below) which the observed value of the statistic cannot be explained by the null model. As with any statistical test, it is necessary to specify a significance level α , which represents the acceptability of rejecting the null model just by chance when it is true. Roughly, values of Tajima's D are significant at the 5% level ($\alpha = 0.05$) if they are either greater than two or less than negative two. However, D is not exactly beta-distributed and critical values are often determined using computer simulation. (This is any area of on-going research.) There are several other related tests that you will probably encounter that are based on the same idea (Fu and Li's D^* and F tests, e.g.).

As far as how the D value responds to deviations from the neutral model, which is the most important thing, this can be understood in the following way. First, the sign of the test is determined only by the sign of the numerator, since the denominator is always positive. The D value becomes negative when there is an excess of either low-frequency (rare) or high-frequency polymorphisms and a deficiency of middle-frequency polymorphisms. This might be caused by positive selection, or, alternatively, expanding population size (note that the Tajima model assumes constant population size for the null hypothesis). Large positive values of D can result from population contraction, or the balancing selection of two alternative polymorphisms. The sensitivity to demographic parameters cannot be overstressed. (Below we turn to a test for selection that does *not* make any such demographic assumptions, the McDonald-Kreitman test; however, it is correspondingly less powerful.)

4. Testing selection vs. neutrality: K_A/K_S ; McDonald-Kreitman (MK) test

Recall from the redundancy of the genetic code that certain nucleotide changes have no effect on the corresponding amino acid coded for – these are called *synonymous* nucleotide substitutions. Otherwise, a substitution is *nonsynonymous* (For example, both CAA and CAG code for glutamine, but CGA codes for arginine, so the first one-letter change alters the amino acid coded for, while the second does not.)

The MK test compares polymorphic and fixed differences found at synonymous and nonsynonymous sites. Because synonymous and nonsynonymous sites are interleaved, one can assume they have the same mutation rate, and so (by taking ratios), we can factor out this usually unknown rate. So, we can test whether the ratio of polymorphism (within species differences) to divergence (between species differences) is the same for both synonymous and nonsynonymous sites. Call K_A the nonsynonymous fixed differences (the “A” reminding us that the change alters the coded-for amino acid), and K_S the synonymous changes. Similarly, within a species, using S for a segregating site as before, we have S_A and S_S . If the neutral theory holds, then $K_A/K_S = S_A/S_S$.

Here’s how to use it.

Consider the evolution of a protein coding gene in two closely related species. Suppose a sample was taken from each of the species. When the sequences from these two samples or populations are aligned together, polymorphic (variable) nucleotide sites can be identified. Each polymorphic site can be classified by two criteria. One is whether the polymorphic site is a difference between samples or a difference between sequences within a sample. Another criteria is whether the change is synonymous. A change is synonymous if it leads to a synonymous codon and otherwise non-synonymous. The result is conveniently presented by the following four values:

	<u>Within species</u>	<u>Between species</u>
Synonymous	a	b
Non-Synonymous	c	d

where a , for example, is the number of polymorphic sites that are both within sample variation and synonymous change. When mutations are selectively neutral, one can expect that the ratio of synonymous and nonsynonymous changes remains constant over time. Therefore, whether a mutation is synonymous should not depend on if it is a within sample polymorphism (occurred recently) or a between sample polymorphism (occurred long time ago). In statistical terms, the two classifications of polymorphic sites are independent under the null hypothesis that mutations are selectively neutral. A simple test of the null hypothesis is a Chi-square test, which is

$$X^2 = n(ad-bc)^2 / [(a+b)(a+c)(b+d)(c+d)]$$

where $n=a+b+c+d$ is the total number of polymorphic sites. When n is not small, X^2 follows approximately a Chi-square distribution with one degree of freedom. So if the value of X^2 is larger than 3.841, the null hypothesis can be rejected at 5% significance level.