Neutral mutations in ideal populations

Gil McVean

Why do we need a mathematical model of evolution?
Population genetics is essentially a mathematical theory of evolution. It aims to take the theory of Darwin and the observations of Mendel and produce a quantitative description of how evolution may proceed in natural populations. In order to achieve this, it is necessary to make gross simplifications of biological reality, and to many empiricists, this is a critical failing of the entire field. What good is it making quantitative predictions when the underlying models contain none of the messiness that makes biology so fascinating to study?

This is a valid concern, and one which needs to be addressed before going further. There are two answers to the criticism. First, no population geneticist believes that the models they use are accurate descriptions of biological reality. But the hope is that the models capture the essence of the problem, and that the vast majority of biological realism is irrelevant to understanding the evolutionary forces responsible for the generation and maintenance of genetic variation. The justification for such optimism is nothing less than Darwin’s theory of evolution by natural selection. Beetles and bacteria are self-evidently very different biological entities, but the very basis of Darwinism is that the same types of evolutionary processes are acting on both.

The second answer is that population genetics, though simplistic, is falsifiable. In fact, it is precisely because of its simplicity that it can be used to make predictions and test hypotheses. For example, suppose we have been travelling the length and breadth of Britain collecting data on genetic variability in beetles, we may want to ask a very simple question such as “does the beetle population show evidence for a recent increase in population size?” If we tried to make a model that included details of how beetles choose a mate, what they eat and who they get eaten by, we would be left with such a complex model that it would be impossible to answer the question we are interested in. But if we start with the simplest possible model – that is we pretend beetles are beans in a beanbag – we can make quantitative predictions about the patterns of variability we expect to see in nature, and test whether the data fit these predictions. By adding extra parameters we can begin to test our assumption of beetle-bean equivalence, and the robustness of our conclusions.

The null model in population genetics
This lecture is about THE null model in population genetics. I want to ask the following question:

Let us suppose that nothing interesting ever happens in biology, and that the amazing complexity of biological populations can be reduced to studying the behaviour of beans in a beanbag; what do we expect patterns of genetic variability to look like?

Which naturally leads to the question:
Do natural populations look like this?

The null model has two key elements – the way in which individuals reproduce, and the way in which new mutations enter the population through mutation. For the moment though, we are going to ignore new mutation, and concentrate on those alleles already segregating in the population.

The ideal population

The most widely used population genetic model for how organisms reproduce is called the Fisher-Wright model, after the two theoreticians who were the first to thoroughly explore its properties. The model has many assumptions, almost none of which are true for any single biological population.

- Constant diploid population of size \(N\) (\(2N\) alleles)
- Non-overlapping generations
- Random mating with respect to both space and genotype
- Hermaphroditism with the possibility of selfing
- No migration to or from other populations
- No selection

Reproduction works by randomly selecting a gamete from one individual, and then randomly selecting another gamete from a second individual (which may be the same as the first) to make one offspring. This is repeated \(N\) times, until a daughter population of \(N\) individuals has been generated. It is easy to see why this type of model is called a beanbag model, because its properties are exactly reproduced by having a bag with \(2N\) beans (some of which may be a different colour if there are several alleles in the population), and picking \(2N\) beans with replacement in order to make a daughter population.

Those with some mathematical background will recognise this model as the binomial distribution. If we sample with replacement a population consisting of \(i\) red beans and \(2N-i\) black beans, the probability of picking \(j\) red beans is

\[
\Pr\{j \mid i\} = \frac{2N!}{j!(2N-j)!} \left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j}
\]

The vertical line in \(\Pr\{j \mid i\}\) means it is a conditional probability – the probability of getting \(j\) in the next generation given that we currently have \(i\). One other point, is that rather than talk about the absolute numbers of alleles in a populations, we usually talk about allele frequencies, and I will use \(x\) to mean \(i/2N\). We can now begin to ask some questions about the behaviour of alleles under such a model.

The relationship between allele frequency and genotype frequency

Because we are dealing with a diploid population, two alleles are found in every individual, and this combination is called the genotype. Under our naive model, what do we expect the genotype frequencies to be?

This problem was first addressed by George Hardy and Wilhelm Weinberg, and the answer has become known as the Hardy-Weinberg law. In fact, they considered a slightly different situation in which the population is so large that it can be treated as an infinite population. When populations are very large, stochastic effects can be ignored. Under these conditions, let us consider the case of two
alleles, A and a, where A is at frequency x. To make a new offspring we draw alleles from the existing population at random, without any regards to the parent from which they have come. So the probability of drawing type A is x, and the probability of drawing one of type a is 1-x. So the probability of drawing each genotype is

\[
\begin{align*}
AA & \quad x^2 \\
Aa & \quad x(1-x) \\
aA & \quad (1-x)x \\
aa & \quad (1-x)^2
\end{align*}
\]

and because the two heterozygotes are equivalent, the total probability of drawing a heterozygote is 2x(1-x). If the population is so large that we can ignore stochastic effects, the expected genotype frequencies in the next generation are equal to the probabilities for every individual drawn. In fact, populations have to be really small in order for stochastic forces to be important in creating deviations from Hardy-Weinberg equilibrium, of the order of 10s rather than 1000s. Hardy Weinberg frequencies are often represented as a Punnett square.

The Hardy-Weinberg law is fairly trivial but it has two key properties. First, allele frequencies are unaffected by the segregation of alleles into different genotypes. Second, Hardy-Weinberg proportions are achieved in just a single round of random mating. For this reason it is perhaps not too surprising that in many populations there is no evidence for deviation from Hardy-Weinberg genotype frequencies (as assessed by a Chi-squared goodness of fit test). For example, this slide shows genotype frequencies for the MN blood group system among different American populations. While there is evidence for differences in allele frequencies between populations, within each one, the Hardy-Weinberg law seems to apply. So in this case, the answer to our second question, of whether natural populations behave like the simple models of population genetics, is an emphatic yes.

Another application of the Hardy-Weinberg law is that we can use phenotypes to estimate underlying genotypes, even when there is dominance or recessivity of the traits involved. For example, cystic fibrosis is an autosomal recessive disease that affects about 1 in 2500 people in Britain. We can estimate the underlying frequency of the disease alleles, and the frequency of carriers by using Hardy-Weinberg proportions.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype</th>
<th>HW frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Normal</td>
<td>(x^2)</td>
</tr>
<tr>
<td>Aa</td>
<td>Normal</td>
<td>(2x(1-x))</td>
</tr>
<tr>
<td>aa</td>
<td>Disease</td>
<td>(x^2)</td>
</tr>
</tbody>
</table>

So if 1 in 2500 people are affected, we solve \(x^2 = 1/2500\) to give \(x = 0.02\), and the frequency of carriers is \(2 \times 0.02 \times 0.98 = 0.04\). In other words, 4% of people carry one copy of a disease allele for cystic fibrosis.
Changes in allele frequency due to genetic drift

The next thing we want to investigate in our null model is how allele frequencies change over time. I have just said that one of the key features of the Hardy-Weinberg law is that allele frequencies are unchanged over time, but in finite populations, allele frequencies fluctuate due to the sampling nature of reproduction, a phenomenon known as genetic drift. For example, if we have 50 copies of an allele in a population of \(2N = 100\), then there is a high chance that we will not get 50 copies by random sampling in the next generation. In fact the probability of getting exactly 50 copies in the next generation is only 0.08. We may get 49 or 58, or in fact any number between 0 and 100.

The binomial distribution characterises the probability of each different outcome. Over a single generation, the expected number of alleles (the sum of all outcomes multiplied by the probability of its occurrence) is unchanged, but there is considerable variation around the expectation. What happens in the next generation? Suppose we had ended up with 43 red alleles in our first daughter population, we would randomly choose another \(2N\) alleles, and in the spirit of the binomial distribution, we may end up with 40, 30 or 75 red alleles in the next population (but the average across an infinite number of trials would be 43). And by combining the distributions derived from every possible outcome of the first round of replication, we can make a distribution of allele frequencies after two rounds of replication. The top chart shows the outcomes for a set of 10 simulated populations, each consisting of 50 diploid individuals.

We can carry on doing this for more and more generations, building up the frequency distribution for allele frequencies after a given period of time. After a while, in some populations, the allele frequency may get so low that in the next generation, none are picked. Alternatively, allele frequency may get so high (approaching 1) that all alleles picked are of that type. Either event is called fixation. Because we are ignoring mutation for the moment, fixation is irreversible, and the frequencies 0 and 1 are called absorbing boundaries. A point worth mentioning because it is highly counter-intuitive, is that after a period of time the frequency distribution almost becomes flat. In other words, across those populations where the allele is still segregating, it is equally likely to be at any frequency, a property that was first noted by Wright in 1931.

What should we make of this distribution, apart from its rather pleasing appearance? Although Wright studied the process intensively, it wasn’t until Kimura (1955) that a full analytical solution to the distribution was available. And even then, it involves such horrendous mathematics that it is almost easier to solve numerically by stochastic simulation than calculate exactly. Yet in the midst of such complex behaviour, one result of quite incredible simplicity emerges.

The decay of heterozygosity due to drift

The remarkable property concerns the quantity we came across in the previous lecture – heterozygosity. Heterozygosity is defined as the proportion of individuals that have two different alleles at a given locus. Under our null model, and with the help of the Hardy-Weinberg law, in our null model the expected frequency of heterozygotes is

\[
H = 2x(1-x)
\]
where $x$ is the allele frequency. Due to stochastic sampling, the actual frequency of heterozygotes may deviate from this, but we are interested in the expected heterozygosity, or the related quantity of average pairwise difference which is the sum across sites in the gene

$$\pi = \left(\frac{2N}{2N-1}\right) \sum_i H_i$$

The remarkable property of the distribution of allele frequencies over time is that whatever the allele frequency, the expected heterozygosity in the next generation is

$$E[H_1] = \sum_{x'} 2x'(1-x') \Pr\{x'|x\} = H_0 \left(1 - \frac{1}{2N}\right)$$

where $\Pr\{x'|x\}$ means the probability of the allele frequency changing from $x$ to $x'$. The expected heterozygosity after two generations is

$$E[H_2] = \sum_{x'} E[H_2 | x'] \Pr\{x'|x\} = \left(1 - \frac{1}{2N}\right) \sum_{x'} 2x'(1-x') \Pr\{x'|x\} = \left(1 - \frac{1}{2N}\right) E[H_1] = H_0 \left(1 - \frac{1}{2N}\right)^2$$

And more generally,

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

In other words, heterozygosity is expected to decay at a steady rate, where a fraction $1/2N$ of the current heterozygosity is lost each generation. Put another way – the rate of loss of heterozygosity is inversely proportional to the population size. This is an important result because it links two things we can readily estimate – genetic heterozygosity and population size.

By now you may be wondering. Where did this come from? What is it about heterozygosity that means it behaves in such a simple manner? There is another way of deriving this result (actually a very closely related result), which is more intuitive, and reveals the connection between the biology and the maths. But I wanted to introduce the subject in this manner, because I think it shows you that heterozygosity is a natural measure to use given the result, rather than assuming that heterozygosity is a natural measure and then deriving the result.

**Identity by descent**

The related derivation uses the notion of identity by descent - the probability that two alleles drawn at random from a population (without replacement) share a common ancestor. In our simple population model everyone is ultimately related to everyone else - everyone in this room shares a common ancestor somewhere. But for the moment, we’ll forget that, and just think about recently shared ancestors.

In our model, the probability that two alleles picked at random in a population are derived from the same allele in the proceeding generation is $1/2N$, so the probability that two alleles are identical by descent in this generation is

$$G_1 = \frac{1}{2N} + (1 - \frac{1}{2N})G_0$$
Let us now consider the related quantity

\[ H_1 = 1 - G_1 \]

\( H \) is the probability that two alleles picked at random are not identical by descent. Following through from the previous equation

\[ H_1 = 1 - G_1 \\
= (1 - \frac{1}{2N})(1 - G_0) \\
= (1 - \frac{1}{2N})H_0 \]

What happens after another round of mating?

\[ G_2 = \frac{1}{2N} + (1 - \frac{1}{2N})G_1 \\
H_2 = (1 - \frac{1}{2N})H_1 \\
= (1 - \frac{1}{2N})^2 H_0 \]

This should be looking familiar by now. In fact, we end up with almost the same relationship as before

\[ H_2 = (1 - \frac{1}{2N})^2 H_0 \]

In other words, through the process of random sampling in a finite population, identity by descent increases such that its complement (the probability that two randomly chosen alleles differ in ancestry) decreases in a steady fashion, at a rate inversely proportional to the population size.

**How much variation? – Identity in state**

We are now in a position to derive our first really important result in population genetics. What I have just outlined is how a simple statistic of polymorphism –heterozygosity – decays in a finite population due to random sampling. But of course, in a real population new mutations are continuously entering the population. So the loss of heterozygosity will eventually come to be balanced by the generation of new heterozygosity.

Where is that equilibrium? We can get the answer by a very subtle change in the previous argument. Rather than talk about identity by descent, I now want to consider the property of identity by state - the probability that two alleles picked at random from the population (without replacement) are different from each other.

\[ G_{t+1} = (1 - u)^2 \left[ \frac{1}{2N} + (1 - \frac{1}{2N})G_t \right] \]

This is almost identical to the beginning of the previous argument, but with the additional of term involving mutation. Essentially the probability of identity in state is the probability that the two alleles picked were identical in state in the previous generation multiplied by the probability that neither allele has mutated in the period since.

At equilibrium, we expect identity in state to be constant over time, so the two sides of the previous equation must be equal to one another. With a little bit of algebraic rearrangement, and ignoring terms involving the square of the mutation rate (we assume mutations are rare, so the probability of more than one mutation in a single generation is almost zero) and also that \( \mu \ll N \) this allows us to solve for \( G \) at equilibrium (the squiggle on top means it is an equilibrium value).
\[ \tilde{G} = \frac{1}{1 + 4Nu} \]

and the related
\[ \tilde{H} = 1 - \tilde{G} = \frac{4Nu}{1 + 4Nu} \]  

(1)

Let us think about this for a second. What I am says is that despite the horrendously complex dynamics of allele frequencies due to mutation and sampling in a finite population, there is a quantity that is both easy to measure, and that captures a lot about levels of polymorphism, that can be predicted in a one line formula. In fact, we can go further, because you will notice that the population size and the mutation rate appear as a product—in general they are fairly inseparable—and the abbreviation

\[ 4Nu \]

is often used. So the expected heterozygosity depends solely on one, compound parameter. In the case of the infinite sites model, we get a very similar result— the expected number of differences between two sequences picked at random from the population is

\[ E[\pi] = 4Nu \]

**An obsession**

It will come as no surprise that this result has created something of an obsession in population genetics. Those new to the field often wonder why there is so much excitement in what is nothing more than a summary statistic. And to be sure, by focusing on heterozygosity alone, you will be ignoring a lot of information in the data. But no other statistic, or characteristic of the data has such a neat relationship to the underlying parameters.

What can we do with this result? Our immediate reaction might be to say that if we can estimate the mutation rate, then we can use the result to estimate the population size for our favourite species (assuming that it behaves as a nice Fisher-Wright population). Let us take for comparison *D. melanogaster* and humans. From protein studies, the average heterozygosities are about 0.15 and 0.06 respectively (there is no evidence for deviation from Hardy-Weinberg equilibrium). What is the rate of mutation to new allozyme variants per locus per generation? This is difficult to estimate accurately, but we can get a rough estimate from levels of protein divergence using the neutral theory result that the rate of neutral substitution is equal to the rate of neutral mutation. For example, the divergence between human and chimp protein sequences is about 0.5%, and they are thought to have diverged about 5 million years ago. If we allow 20 years per generation in the human lineage, we get an estimate of \(10^8\) mutations per amino acid per generation. There are about 500 amino acids per gene, so the rate of change per gene is about \(5 \times 10^6\) per generation. If 50% of all amino acid changes lead to new electrophoretic variants, then the neutral rate of mutation to new allozymes is about \(2.5 \times 10^6\) per generation. Solving for \(N\) in our simple equation gives 6,000. Doing a similar thing for *Drosophila* provides an estimate of about 200,000.

There is a problem. *D. melanogaster* is a cosmopolitan species, it occurs almost everywhere in the world, and almost everywhere it occurs, it occurs in large numbers. Its current population is
probably of the order of 100s of billions, not 100s of thousands. The same can be said for humans, though on an evolutionary time scale, their numbers must have been much less – but as few as 6,000? What’s gone wrong? Perhaps some of the numbers fed in to the neutral mutation rate estimate are incorrect? Almost certainly, but they are not out by orders of magnitude. And you don’t even need these numbers to spot the problem. Humans have about half the protein heterozygosity of *Drosophila*, but their population size must be many thousand fold smaller than that of *Drosophila*.

**Effective population size**

However you tweak the numbers that go into estimating population size directly from equation (1), it is impossible to come up with biologically plausible of estimates of population size directly from levels of heterozygosity. Clearly, our naïve beanbag model of alleles in populations is wrong. But the difficult question, is where to go from here? We have no desire to reject the model outright, because it leaves us with no alternative framework for interpreting patterns of genetic variability. Can we find some way of keeping the simplicity of the model structure, but introducing important biological complications?

The solution is so subtle that it is almost semantic. It turns out that the effect of many biological complications on genetic variability, such as variation in reproductive success, fluctuations in population size over time, and differences in the number of breeding males and females, is to make to population behave as if it were still our naïve Fisher-Wright model, but one in which the population size is smaller than the true census population size. This transformed population size is called the effective population size and is written as $N_e$.

It is worth going through some examples of how the effective population size can be calculated to get a feel for how different $N_e$ can be from $N$.

*Variation in breeding success.* Suppose that our model for how alleles are chosen for the next generation is incorrect, and the binomial sampling procedure underestimates the true variation in breeding success. For example, our population may be constant over time, but a large fraction are killed off each year by random factors (earthquakes and avalanches). How does this affect genetic variability? As before, we work out the change per generation in the probability of identity by descent. If the number of offspring left by allele $i$ is $k$, the probability that two alleles drawn at random from the population came from the same parent in the previous generation is

$$
\text{Pr\{same parent\}} = \sum_{i} \frac{k_i(k_i - 1)}{2N(2N - 1)}
$$

$$
= \frac{1}{2N - 1} \left[ \frac{1}{2N} \sum_i k_i^2 - \frac{1}{2N} \sum_i k_i \right]
$$

$$
= \frac{1}{2N - 1} \left[ \sigma_k^2 + \bar{k}^2 - \bar{k} \right]
$$

$$
\approx \frac{\sigma_k^2}{2N} \text{ if } \bar{k} = 1
$$
So if the total population is constant, the probability of two alleles being different by descent increases as

\[ H_{t+1} = H_t (1 - \frac{\sigma_k^2}{2N}) \]

If we then equate the rate of change in identity with that expected under some smaller Fisher Wright population of size \( N_e \)

\[ H_{t+1} = H_t (1 - \frac{1}{2N_e}) \]

it follows that

\[ N_e = \frac{N}{\sigma_k^2} \]

Under the binomial sampling model, the variance in offspring number is approximately one for large \( N \), so \( N_e \) is equal to \( N \).

**Fluctuation in population size.** Random variation in population size over time can also be treated as a reduction in the effective population size. Again, consider the rate at which heterozygosity decays in the population.

\[ H_1 = H_0 (1 - \frac{1}{2N_0}) \]
\[ H_2 = H_1 (1 - \frac{1}{2N_1})H_1 \]
\[ = H_0 (1 - \frac{1}{2N_0})(1 - \frac{1}{2N_1}) \]
\[ H_{t+1} = H_0 (1 - \frac{1}{2N_0})(1 - \frac{1}{2N_1})... (1 - \frac{1}{2N_t}) \]
\[ = H_0 \prod_{i}(1 - \frac{1}{2N_i}) \]

In order to find a solution for \( N_e \), we need to make use of the behaviour of exponentials when \( N >> 1 \).

\[ (1 - \frac{1}{2N}) \approx e^{-1/2N} \]
\[ \prod_{i}(1 - \frac{1}{2N_i}) \approx e^{\sum_{i} \frac{1}{2N_i}} \]

So by equating the sum with the equivalent term involving \( N_e \)

\[ e^{\sum_{i} \frac{1}{2N_i}} = e^{\frac{t}{2N_e}} \]
\[ \Rightarrow N_e = t \sum_{i} \frac{1}{2N_i} \]

In other words, when there is random fluctuation in population size, the effective population size is the harmonic mean of the population sizes. The harmonic mean is strongly influenced by small values, so populations which are usually large, but occasionally collapse to much smaller ones for brief periods can have quite small effective population sizes.

**Differences in the numbers of breeding males and females.** One final case is worth noting. In many mammalian species, strong sexual selection means that there is a great skew in the reproductive success...
of males and females. In species such as red deer and elephant seals, where a single male controls a
group of females, the variance in breeding success will be very different between the sexes. This type
of population can have an effective population size much lower than the census size. The formula
relating the number of breeding males and females to the effective population size is

\[ N_e = \frac{4N_mN_f}{N_m + N_f} \]

In the example given, if only 1/6 of all males actually mate, the effective population size is about 30% of the census size.

Other biological complexities can also be treated as a change in effective population size, and there is a
whole industry in theoretical population genetics of deriving formulas for \( N_e \) under various
circumstances. There is no need to remember formulas, just two consequences of this shift in thinking.
First, the effective population size has very little to do with the actual, or census population size. It is
almost always smaller than the census size, and often very considerably so. The second point is that
we can reconcile our problems to do with the constancy of levels of genetic variability across natural
populations, if we say that while their census populations may vary enormously, their effective
population sizes do not.

**Algebraic backhand or deep insight?**

This neat piece of algebraic problem solving should hopefully leave you feeling a little uneasy. I have
just said that we can explain away inconsistencies in applying the Fisher-Wright model to real
populations by saying that levels of variability are the result of a number, \( N_e \), that we cannot hope to
measure.

We need a way of testing whether our transformation of \( N \) into \( N_e \) is justified. And to do that
we need to move away from our favourite statistic, heterozygosity, in order to look at other features of
genetic variability – such as how many alleles we would expect to find in a population, and what their
frequency distribution should look like. In addition, we need a shift in perspective - from the process
of parameter estimation, to that of hypothesis testing.

**Diffusion theory**

In order to get any further with the Fisher-Wright model it turns out that we need to introduce a new
way of thinking about alleles in populations. Until now, we have been dealing with discrete alleles,
represented by an integer, \( i \) in a population of \( N \) individuals. However, as \( N \) gets really large, the
difference in allele frequency \( (i/2N) \) between \( i \) copies and \( i+1 \) copies gets so small, that if we blur our
eyes a bit, we can pretend that allele frequency is actually continuous.

How does this help? By pretending allele frequency is continuous rather than discrete we can
make use of mathematical techniques originally developed in physics by the Russian Kolmogorov for
describing how particles move in space. These methods, collectively known as diffusion theory, have a
truly great history, and have helped solve some key problems in physics. In genetics, diffusion theory
has been used to solve problems relating to the distribution of allele frequencies – for example the distributions we solved numerically earlier on in the lecture.

Although the algebra looks nasty at times, the underlying idea behind diffusion theory is fairly intuitive – that by looking at how systems change over very short periods of time, we can predict their long term behaviour. In terms of genetics, we have to assume

A) Changes in allele frequency each generation due to drift, mutation, migration or selection are small.

B) The only source of variation in change in allele frequency each generation is that due to random (binomial) sampling.

C) The mean and variance of allele frequency change are sufficient to fully describe the population.

Wright was the first to develop the use of diffusion methods in population genetics. Kimura subsequently solved many of the thornier problems. The important point to note is that diffusion methods allow us to derive results concerning the distribution of allele frequencies in populations.

Exact solutions to these distributions in our discrete allele, Fisher-Wright model are not available.

**Derivation of the forward Kolmogorov equation**

This section is not critical to the lecture, but for those interested, I have included a derivation of the forward Kolmogorov equation – which has been used to investigate the distribution of allele frequencies in populations.

Think back to our original simulations where we followed the fate of alleles in a collection of 10 replicate populations. Now let us suppose that we have an infinite collection of such populations. What would be the distribution of the proportion of populations in which there were $i$ copies of the allele after $t$ generations? Let us suppose that at $t-1$ generations, the proportion of populations in which there were $i$ copies was $f(i)$. We can write the proportion of populations in which there are $i$ copies in the next generation as a sum over all previous generations

$$f(i)_t = \sum_{\Delta} f(i+\Delta)_{t-1} \Pr[\Delta]$$

(2)

where $\Pr[\Delta]$ means the probability of going from the state of having $i+\Delta$ copies to having $i$ copies (the binomial transition probability from earlier). Let us now make the switch to continuous allele frequencies, and introduce the notion $\phi(x;t)$ to mean the density of populations at time $t$ in which the allele frequency is $x$ ($\phi(x;t)\delta x$ is the proportion of populations in which the allele frequency is in the range $x$ to $x+\delta x$). We also want to switch to a continuous representation of time. With this in mind, let us also write

$$g(x-\epsilon,\epsilon,\delta t)$$

to mean the probability of going from allele frequency $x-\epsilon$ to $x$ in time $\delta t$. Replacing the summation (2) with an integral, gives something almost identical in spirit

$$\phi(x; t + \delta t) = \int \phi(x-\epsilon; t) g(x-\epsilon, \epsilon, \delta t) d\epsilon$$

(3)
(The critical difference is the $\phi$ is a probability density function whereas $f$ actually represents a probability). The reason for converting to continuous allele frequency and time now becomes clear, because we can make use a Taylor expansion to approximate the product on the RHS of equation (3).

$$\phi(x-\epsilon,t) g(x-\epsilon,\epsilon,\delta t) = \phi g - \epsilon \frac{\partial(\phi g)}{\partial x} + \frac{\epsilon^2}{2} \frac{\partial^2(\phi g)}{\partial x^2} + o(\epsilon^3)$$

Where the function $\phi g$ and its derivatives are evaluated at $\epsilon = 0$. Changing the order between integration, summation and differentiation, and neglecting the terms in $\epsilon^3$, gives

$$\phi(x,t+\delta t) = \phi \left[ g d\epsilon - \frac{\partial}{\partial x} \left\{ \phi \left[ \epsilon g \; d\epsilon \right] \right\} + \frac{\partial^2}{\partial x^2} \left\{ \phi \left[ \epsilon^2 g \; d\epsilon \right] \right\} \right]$$

Note that the integral in the first term is equivalent to the sum of all possible transitions, and must therefore equal one. We can then transfer the first term to the LHS and divide through by $\delta t$ to give

$$\frac{\phi(x,t+\delta t) - \phi(x,t)}{\delta t} = -\frac{\partial}{\partial x} \left\{ \phi(x,t) \frac{1}{\delta t} \left[ \epsilon g(x-\epsilon,\epsilon,\delta t) d\epsilon \right] \right\} + \frac{\partial^2}{\partial x^2} \left\{ \phi(x,t) \frac{1}{\delta t} \left[ \epsilon^2 g(x-\epsilon,\epsilon,\delta t) d\epsilon \right] \right\}$$

The integral in the first term on the RHS is equivalent to expected change in allele frequency in the time $\delta t$ given $x$. The second term is the expectation of the square of the change in allele frequency. Because

$$\sigma_y^2 = E[y^2] - E[y]^2$$

if we say that the expected change in allele frequency is very small, the expectation of the square is approximately the variance in allele frequency change, which is the binomial sampling variance. So, if we define

$$M(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \left[ \epsilon g(x-\epsilon,\epsilon,\delta t) d\epsilon \right]$$

$$V(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \left[ \epsilon^2 g(x-\epsilon,\epsilon,\delta t) d\epsilon \right]$$

then we derive the forward Kolmogorov equation

$$\frac{\partial \phi(x,t)}{\partial t} = \frac{1}{2} \frac{\partial^2}{\partial x^2} \left\{ V(x,t) \phi(x,t) \right\} - \frac{\partial}{\partial x} \left\{ M(x,t) \phi(x,t) \right\}$$

In short, we have found a way of describing how the probability density function behaves in terms of nothing more than the mean and variance of allele frequency change. At equilibrium, the rate of change of the function is zero, which gives us a means of solving the exact form of the stationary distribution. Crow and Kimura (1970) write this in a very neat form

$$\phi(x) = \frac{C}{V(x,t)} \exp \left[ 2 \int \frac{M(x,t)}{V(x,t)} \; dx \right]$$

where $C$ is a normalising constant. A derivation of this result can be found in Crow & Kimura (1970).

**The frequency distribution in the infinite alleles model**

As an example of the application of the diffusion result, consider the infinite alleles model. The expected change in allele frequency per generation is

Copyright: Gilean McVean, 2001
\[ M(x,t) = -ux \]

because mutations to new alleles occur at the rate \( u \). The variance in allele frequency change is

\[ V(x,t) = \frac{x(1-x)}{2N_e} \]

so

\[ 2 \frac{M(x,t)}{V(x,t)} = \frac{4N_e u}{1-x} \]

and at equilibrium

\[ \phi(x) = \frac{2N_e C}{x(1-x)} (1-x)^{4N_e u} \]

Solving for \( C \) gives

\[ \phi(x) = \theta x^{-1} (1-x)^{\theta^{-1}} \text{ where } \theta = 4N_e u \]

An example of this distribution is shown with \( \theta = 0.1 \). The important thing to note is that most alleles exist at low frequency.

**Ewens sampling theory**

Diffusion theory methods can be used to find a solution to the distribution of allele frequencies across our infinite set of identical populations. This is a very important step, because it means we can begin to test our effective population size model for genetic variation by comparing observed allele frequency distributions with those predicted under the neutral model.

But there is one last hurdle we need to get over before we can actually test real data. The problem is that when we sample alleles from populations, we are not directly observing allele frequencies, we are only estimating them. And the difference between a sample and the population can be considerable. For example, it turns out that under the infinite alleles model, the expected number of alleles in the entire population approaches infinity as the population approaches infinity. But the vast majority of these are at such low frequency, that your chances of picking them up in a random sample of 20 or 30 alleles are negligible.

The first person to make the shift in focus from populations to samples was the Australian Warren Ewens. In 1972, he published a paper on the expected number of alleles in a sample under the infinite alleles model. He derived the result that the expected number of alleles in a sample of size \( n \) is

\[ E[K] = 1 + \frac{\theta}{1 + \theta} + \frac{\theta}{2 + \theta} + \ldots + \frac{\theta}{n-1+\theta} \]

This is a remarkable result for two reasons. First, it depends only on the parameter \( n \) and the product \( 4N_e u \), the census population size does not enter the equation at all. The second thing to note is that as \( n \) increases, the expected number of alleles in the sample increases at a diminishing rate. For example, if \( q = 0.5 \), you would expect to find 2.1 different alleles in a sample of size 10, and 2.5 alleles in a sample of size 20.
In the same paper, Ewens produced another remarkable result. He showed that under the neutral, infinite alleles model, the probability of a particular allele configuration, given that \( k \) alleles have been observed is

\[
\Pr\{n_1, n_2, \ldots, n_k \mid k, n\} = \frac{n!}{k! l_k n_1 n_2 \ldots n_k}
\]

where \( n_i \) is the number of copies of allele \( i \) in the sample of size \( n \), and \( l_k \) is normalising constant (for what it’s worth it is a Stirling’s number of the first kind). This result is of huge importance, because it means that we have for the first time a way of testing the hypothesis that out sample of alleles conforms to the effective population size model. By simulating samples from Ewens’ formula, we can look at whether the frequency distribution of alleles that we observe is compatible with the standard neutral model.

In order to do this, we need a way of summarising information about the frequency spectrum. One possibility, suggested by Ewens, is to use a measure of how even the allele frequency distribution is. In particular, we can use the statistic

\[
B = \sum_i x_i \ln(x_i)
\]

where \( x_i \) is the frequency of allele \( i \) in our sample. When the allele frequency is even, \( B \) is a large negative number. In contrast, if there is one major allele, and several minor alleles (typically what we expect under the neutral model \( B \) will be a small negative number. By generating several thousand samples under the null model, we can ask whether our observed value of \( B \) (in this case –2.29 from a sample of 228 individuals at the HLA-A locus in European caucasoids, with a total of 15 alleles) is more negative than we would expect by chance. It turns out in this case that we can reject the hypothesis that our data conform to the standard neutral model at the 4% level. With a bit of relief, we can say that interesting things do, in fact, happen in biology.

**Samples v populations**

Perhaps the most important consequence of Ewens’ paper was that it shifted the focus from studies of whole populations to studies of samples from populations. In the next lecture, I will introduce a way of thinking about samples from populations that is completely changing the way inference in population genetics is carried out. It is called the coalescent.