Genetic Drift

Introduction

So far in this course we've talked about changes in genotype and allele frequencies as if they were completely deterministic. Given the current allele frequencies and viabilities, for example, we wrote down an equation describing how they will change from one generation to the next:

$$p' = \frac{p^2 w_{11} + pq w_{12}}{\bar{w}} \quad .$$

Notice that in writing this equation, we're claiming that we can predict the allele frequency in the next generation *without error*. But suppose the population is small, say 10 diploid individuals, and our prediction is that $p' = 0.5$. Then just as we wouldn't be surprised if we flipped a coin 20 times and got 12 heads, we shouldn't be surprised if we found that $p' = 0.6$. The difference between what we expect ($p' = 0.5$) and what we observe ($p' = 0.6$) can be chalked up to statistical sampling error. That sampling error is the cause of (or just another name for) *genetic drift*—the tendency for allele frequencies to change from one generation to the next in a finite population even if there is no selection.

A simple example

To understand in more detail what happens when there is genetic drift, let's consider the simplest possible example: a haploid population consisting of 2 individuals.[1] Suppose that we are studying a locus with only two alleles in this population A_1 and A_2. This implies that $p = q = 0.5$, but we'll ignore that numerical fact for now and simply label the frequency of the A_1 allele as p.

We imagine the following scenario:

- Each individual in the population produces a very large number of haploid gametes that develop directly into adult offspring.

[1] Notice that once we start talking about genetic drift, we have to specify the size of the population.

© 2001-2019 Kent E. Holsinger

- The allele in each offspring is an identical copy of the allele in its parent, i.e., A_1 begets A_1 and A_2 begets A_2. In other words, there's no mutation.
- The next generation is constructed by picking two offspring at random from the very large number of offspring produced by these two individuals.

Then it's not too hard to see that

$$\begin{aligned} \text{Probability that both offspring are } A_1 &= p^2 \\ \text{Probability that one offspring is } A_1 \text{ and one is } A_2 &= 2pq \\ \text{Probability that both offspring are } A_2 &= q^2 \end{aligned}$$

Of course $p' = 1$ if both offspring sampled are A_1, $p' = 1/2$ if one is A_1 and one is A_2, and $p' = 0$ if both are A_2, so that set of equations is equivalent to this one:

$$P(p' = 1) = p^2 \tag{1}$$
$$P(p' = 1/2) = 2pq \tag{2}$$
$$P(p' = 0) = q^2 \tag{3}$$

In other words, we can no longer predict with certainty what allele frequencies in the next generation will be. We can only assign probabilities to each of the three possible outcomes. Of course, in a larger population the amount of uncertainty about the allele frequencies will be smaller,[2] but there will be *some* uncertainty associated with the predicted allele frequencies unless the population is infinite.

The probability of ending up in any of the three possible states obviously depends on the current allele frequency. In probability theory we express this dependence by writing equations (1)–(3) as conditional probabilities:

$$P(p_1 = 1|p_0) = p_0^2 \tag{4}$$
$$P(p_1 = 1/2|p_0) = 2p_0q_0 \tag{5}$$
$$P(p_1 = 0|p_0) = q_0^2 \tag{6}$$

I've introduced the subscripts so that we can distinguish among various generations in the process. Why? Because if we can write equations (4)–(6), we can also write the following equations:[3]

$$\begin{aligned} P(p_2 = 1|p_1) &= p_1^2 \\ P(p_2 = 1/2|p_1) &= 2p_1q_1 \\ P(p_2 = 0|p_1) &= q_1^2 \end{aligned}$$

[2] More about that later.

[3] I know. I'm weird. I actually get a kick out of writing equations!

Now if we stare at those a little while, we[4] begin to see some interesting possibilities. Namely,

$$\begin{aligned}
P(p_2 = 1|p_0) &= P(p_2 = 1|p_1 = 1)P(p_1 = 1|p_0) + P(p_2 = 1|p_1 = 1/2)P(p_1 = 1/2|p_0) \\
&= (1)(p_0^2) + (1/4)(2p_0q_0) \\
&= p_0^2 + (1/2)p_0q_0 \\
P(p_2 = 1/2|p_0) &= P(p_2 = 1/2|p_1 = 1/2)P(p_1 = 1/2|p_0) \\
&= (1/2)(2p_0q_0) \\
&= p_0q_0 \\
P(p_2 = 0|p_0) &= P(p_2 = 0|p_1 = 0)P(p_1 = 0|p_0) + P(p_2 = 0|p_1 = 1/2)P(p_1 = 1/2|p_0) \\
&= (1)(q_0^2) + (1/4)(2p_0q_0) \\
&= q_0^2 + (1/2)p_0q_0
\end{aligned}$$

It takes more algebra than I care to show,[5] but these equations can be extended to an arbitrary number of generations.

$$\begin{aligned}
P(p_t = 1|p_0) &= p_0^2 + \left(1 - (1/2)^{t-1}\right) p_0q_0 \\
P(p_t = 1/2|p_0) &= p_0q_0(1/2)^{t-2} \\
P(p_t = 0|p_0) &= q_0^2 + \left(1 - (1/2)^{t-1}\right) p_0q_0
\end{aligned}$$

Why do I bother to show you these equations?[6] Because you can see pretty quickly that as t gets big, i.e., the longer our population evolves, the smaller the probability that $p_t = 1/2$ becomes. In fact, it's not hard to verify two facts about genetic drift in this simple situation:

1. One of the two alleles originally present in the population is certain to be lost eventually.
2. The probability that A_1 is fixed is equal to its initial frequency, p_0, and the probability that A_2 is fixed is equal to its initial frequency, q_0.

Both of these properties are true in general for *any* finite population and *any* number of alleles.

1. Genetic drift will eventually lead to loss of all alleles in the population except one.[7]
2. The probability that any allele will eventually become fixed in the population is equal to its current frequency.

[4] Or at least the weird ones among us

[5] Ask me, if you're really interested.

[6] It's not just that I'm crazy.

[7] You obviously can't lose all of them unless the population becomes extinct.

General properties of genetic drift

What I've shown you so far applies only to a haploid population with two individuals. Even I will admit that it isn't a very interesting situation. Suppose, however, we now consider a populaton with N diploid individuals. We can treat it as if it were a population of $2N$ haploid individuals using a direct analogy to the process I described earlier, and then things start to get a little more interesting.

- Each individual in the population produces a large number of gametes.
- The allele in each gamete is an identical copy of the allele in the individual that produced it, i.e., A_1 begets A_1 and A_2 begets A_2.
- The next generation is constructed by picking $2N$ gametes at random from the large number originally produced.

We can then write a general expression for how allele frequencies will change between generations. Specifically, the distribution describing the probability that there will be j copies of A_1 in the next generation given that there are i copies in this generation is

$$P(j \text{ } A_1 \text{ in offspring} \mid i \text{ } A_1 \text{ in parents}) = \binom{2N}{j} \left(\frac{i}{2N}\right) \left(1 - \frac{i}{2N}\right) \quad ,$$

i.e., a binomial distribution. I'll be astonished if any of what I'm about to say is apparent to any of you from looking at this equation, but it implies three really important things. We've encountered the first two of them already:

- Allele frequencies will tend to change from one generation to the next purely as a result of sampling error. As a consequence, genetic drift will eventually lead to loss of all alleles in the population except one.
- The probability that any allele will eventually become fixed in the population is equal to its current frequency.
- The population has no memory.[8] The probability that the offspring generation will have a particular allele frequency depends *only* on the allele frequency in the parental generation. It does not depend on how the parental generation came to have that allele frequency. This is exactly analogous to coin-tossing. The probability that you get a

[8]Technically, we've described a Markov chain with a finite state space, but I doubt that you really care about that. All Markov chains have this "memoryless" property. In fact, it's called the Markov property.

heads on the next toss of a fair coin is 1/2. It doesn't matter whether you've never tossed it before or if you've just tossed 25 heads in a row.[9]

Variance of allele frequencies between generations

For a binomial distribution

$$\begin{aligned} P(K=k) &= \binom{N}{k} p^k (1-p)^{N-k} \\ \mathrm{Var}(K) &= Np(1-p) \\ \mathrm{Var}(p) &= \mathrm{Var}(K/N) \\ &= \frac{1}{N^2}\mathrm{Var}(K) \\ &= \frac{p(1-p)}{N} \end{aligned}$$

Applying this to our situation,

$$\mathrm{Var}(p_{t+1}) = \frac{p_t(1-p_t)}{2N}$$

$\mathrm{Var}(p_{t+1})$ measures the amount of uncertainty about allele frequencies in the next generation, given the current allele frequency. As you probably guessed long ago, the amount of uncertainty is inversely proportional to population size. The larger the population, the smaller the uncertainty.

If you think about this a bit, you might expect that a smaller variance would "slow down" the process of genetic drift — and you'd be right. It takes some pretty advanced mathematics to say how much the process slows down as a function of population size,[10] but we can summarize the result in the following equation:

$$\bar{t} \approx -4N\left(p\log p + (1-p)\log(1-p)\right) \quad ,$$

where $\bar{t}$ is the average time to fixation of one allele or the other and p is the current allele frequency.[11] So the average time to fixation of one allele or the other increases approximately linearly with increases in the population size.

[9] Of course, if you've just tossed 25 heads in a row, you could be forgiven for having your doubts about whether the coin is actually fair.

[10] Actually, we'll encounter a way that isn't quite so hard in a few lectures when we get to the coalescent.

[11] Notice that this equation only applies to the case of one-locus with two alleles, although the principle applies to any number of alleles.

Analogy to inbreeding

You may have noticed some similarities between drift and inbreeding. Specifically, both processes lead to a loss of heterozygosity and an increase in homozygosity. This analogy leads to a useful heuristic for helping us to understand the dynamics of genetic drift.

Remember our old friend f, the inbreeding coefficient? I'm going to re-introduce you to it in the form of the population inbreeding coefficient, the probability that two alleles chosen at random from a population are identical by descent. We're going to study how the population inbreeding coefficient changes from one generation to the next as a result of reproduction in a finite population.[12]

$$\begin{aligned} f_{t+1} &= \text{Prob. ibd from preceding generation} \\ &\quad +(\text{Prob. not ibd from prec. gen.}) \times (\text{Prob. ibd from earlier gen.}) \\ &= \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t \end{aligned}$$

or, in general,

$$f_{t+1} = 1 - \left(1 - \frac{1}{2N}\right)^t (1 - f_0) \quad .$$

Summary

There are four characteristics of genetic drift that are particularly important for you to remember:

1. Allele frequencies tend to change from one generation to the next simply as a result of sampling error. We can specify a probability distribution for the allele frequency in the next generation, but we cannot predict the actual frequency with certainty.

2. There is no systematic bias to changes in allele frequency. The allele frequency is as likely to increase from one generation to the next as it is to decrease.

3. If the process is allowed to continue long enough without input of new genetic material through migration or mutation, the population will eventually become fixed for only one of the alleles originally present.[13]

[12] Remember that I use the abbreviation ibd to mean identical by descent.

[13] This will hold true even if there is strong selection for keeping alleles in the population. Selection can't prevent loss of diversity, only slow it down.

4. The time to fixation on a single allele is directly proportional to population size, and the amount of uncertainty associated with allele frequencies from one generation to the next is inversely related to population size.

Effective population size

I didn't make a big point of it, but in our discussion of genetic drift so far we've assumed everything about populations that we assumed to derive the Hardy-Weinberg principle, *and* we've assumed that:

- We can model drift in a finite population as a result of sampling among haploid gametes rather than as a result of sampling among diploid genotypes. Since we're dealing with a finite population, this effectively means that the two gametes incorporated into an individual could have come from the same parent, i.e., some amount of self-fertilization can occur when there's random union of gametes in a finite, diploid population.
- Since we're sampling gametes rather than individuals, we're also implictly assuming that there aren't separate sexes.[14]
- The number of gametes any individual has represented in the next generation is a binomial random variable.[15]
- The population size is constant.

How do we deal with the fact that one or more of these conditions will be violated in just about any case we're interested in?[16] One way would be to develop all the probability models that incorporate that complexity and try to solve them. That's nearly impossible, except through computer simulations. Another, and by far the most common approach, is to come up with a conversion formula that makes our actual population seem like the "ideal" population that we've been studying. That's exactly what *effective population size* is.

> The effective size of a population is the size of an ideal population that has the same properties with respect to genetic drift as our actual population does.

What does that phrase "same properties with respect to genetic drift" mean? Well there are two ways it can be defined.[17]

[14] How could there be separate sexes if there can be self-fertilization?
[15] More about this later.
[16] OK, OK. They will probably be violated in *every* case we're interested in.
[17] There are actually more than two ways, but we're only going to talk about two.

Variance effective size

You may remember[18] that the variance in allele frequency in an ideal population is

$$Var(p_{t+1}) = \frac{p_t(1-p_t)}{2N} \quad .$$

So one way we can make our actual population equivalent to an ideal population to make their allele frequency variances the same. We do this by calculating the variance in allele frequency for our actual population, figuring out what size of ideal population would produce the same variance, and pretending that our actual population is the same as an ideal population of the same size. To put that into an equation,[19] let $\widehat{Var}(p)$ be the variance we calculate for our actual population. Then

$$N_e^{(v)} = \frac{p(1-p)}{2\widehat{Var}(p)}$$

is the *variance effective population size*, i.e., the size of an ideal population that has the same properties with respect to allele frequency variance as our actual population.

Inbreeding effective size

You may also remember that we can think of genetic drift as analogous to inbreeding. The probability of identity by descent within populations changes in a predictable way in relation to population size, namely

$$f_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t \quad .$$

So another way we can make our actual population equivalent to an ideal population is to make them equivalent with respect to how f changes from generation to generation. We do this by calculating how the inbreeding coefficient changes from one generation to the next in our actual population, figuring out what size an ideal population would have to be to show the same change between generations, and pretending that our actual population is the same size at the ideal one. So suppose $\hat{f}_t$ and $\hat{f}_{t+1}$ are the actual inbreeding coefficients we'd have in our population at generation t and $t+1$, respectively. Then

$$\begin{aligned} \hat{f}_{t+1} &= \frac{1}{2N_e^{(f)}} + \left(1 - \frac{1}{2N_e^{(f)}}\right) \hat{f}_t \\ &= \left(\frac{1}{2N_e^{(f)}}\right)(1 - \hat{f}_t) + \hat{f}_t \end{aligned}$$

[18] You probably won't, so I'll remind you

[19] As if that will make it any clearer. Does anyone actually read these footnotes?

$$\begin{aligned} \hat{f}_{t+1} - \hat{f}_t &= \left(\frac{1}{2N_e^{(f)}}\right)(1 - \hat{f}_t) \\ N_e^{(f)} &= \frac{1 - \hat{f}_t}{2(\hat{f}_{t+1} - \hat{f}_t)} \quad . \end{aligned}$$

In many applications it's convenient to assume that $\hat{f}_t = 0$. In that case the calculation gets much simpler:

$$N_e^{(f)} = \frac{1}{2\hat{f}_{t+1}} \quad .$$

We also don't lose anything by taking the simpler approach, because $N_e^{(f)}$ depends only on how much f *changes* from one generation to the next, not on its actual magnitude.

Comments on effective population sizes

Those are nice tricks, but there are some limitations. The biggest is that $N_e^{(v)} \neq N_e^{(f)}$ if the population size is changing from one generation to the next.[20] So you have to decide which of these two measures is more appropriate for the question you're studying.

- $N_e^{(f)}$ is naturally related to the number of individuals in the parental populations. It tells you something about how the probability of identity by descent within a single population will change over time.
- $N_e^{(v)}$ is naturally related to the number of individuals in the offspring generation. It tells you something about how much allele frequencies in isolated populations will diverge from one another.

Examples

This is all pretty abstract. Let's work through some examples to see how this all plays out.[21] In the case of separate sexes and variable population size, I'll provide a derivation of $N_e^{(f)}$. In the case of differences in the number of offspring left by individuals, I'll just give you the formula and we'll discuss some of the implications.

[20] It's even worse than that. When the population size is changing, it's not clear that any of the available adjustments to produce an effective population size are entirely satisfactory. Well, that's not entirely true either. Fu et al. [2] show that there is a reasonable definition in one simple case when the population size varies, and it happens to correspond to the solution presented below.

[21] If you're interested in a comprehensive list of formulas relating various demographic parameters to effective population size, take a look at [1, p. 362]. They provide a pretty comprehensive summary and a number of derivations.

Separate sexes

We'll start by assuming that $\hat{f}_t = 0$ to make the calculations simple. So we know that

$$N_e^{(f)} = \frac{1}{2\hat{f}_{t+1}} \quad .$$

The first thing to do is to calculate $\hat{f}_{t+1}$. To do this we have to break the problem down into pieces.[22]

- We assumed that $\hat{f}_t = 0$, so the only way for two alleles to be identical by descent is if they are identical copies of the *same* allele in the immediately preceding generation.
- Even if the numbers of reproductive males and reproductive females are different, every new offspring has exactly one father and one mother. Thus, the probability that the first gamete selected at random is female is just 1/2, and the probability that the first gamete selected is male is just 1/2.
- The probability that the second gamete selected is female given that the first one we selected was female is $(N-1)/(2N-1)$, because N out of the $2N$ alleles represented among offspring came from females, and there are only $N-1$ out of $2N-1$ left after we've already picked one. The same logic applies for male gametes.
- The probability that one particular female gamete was chosen is $1/2N_f$, where N_f is the number of females in the population. Similarly the probability that one particular male gamete was chosen is $1/2N_m$, where N_m is the number of males in the population.

With those facts in hand, we're ready to calculate $\hat{f}_{t+1}$.

$$\begin{aligned}
f_{t+1} &= \left(\frac{1}{2}\right)\left(\frac{N-1}{2N-1}\right)\left(\frac{1}{2N_f}\right) + \left(\frac{1}{2}\right)\left(\frac{N-1}{2N-1}\right)\left(\frac{1}{2N_m}\right) \\
&= \left(\frac{1}{2}\right)\left(\frac{N-1}{2N-1}\right)\left(\frac{1}{2N_f} + \frac{1}{2N_m}\right) \\
&\approx \left(\frac{1}{2}\right)\left(\frac{2N_m + 2N_f}{4N_f N_m}\right) \\
&= \left(\frac{1}{2}\right)\left(\frac{N_m + N_f}{4N_f N_m}\right)
\end{aligned}$$

[22] Remembering, of course, that $\hat{f}_{t+1}$ is the probability that two alleles drawn at random are identical by descent.

So,

$$N_e^{(f)} \approx \frac{4N_f N_m}{N_f + N_m} \quad .$$

What does this all mean? Well, consider a couple of important examples. Suppose the numbers of females and males in a population are equal, $N_f = N_m = N/2$. Then

$$\begin{aligned} N_e^{(f)} &= \frac{4(N/2)(N/2)}{N/2 + N/2} \\ &= \frac{4N^2/4}{N} \\ &= N \quad . \end{aligned}$$

The effective population size is equal to the actual population size if the sex ratio is 50:50. If it departs from 50:50, the effective population size will be smaller than the actual population size. Consider the extreme case where there's only one reproductive male in the population. Then

$$N_e^{(f)} = \frac{4N_f}{N_f + 1} \quad . \tag{7}$$

Notice what this equation implies: The effective size of a population with only one reproductive male (or female) can *never* be bigger than 4, no matter how many mates that individual has and no matter how many offspring are produced.

Variable population size

The notation for this one gets a little more complicated, but the ideas are simpler than those you just survived. Since the population size is changing we need to specify the population size at each time step. Let N_t be the population size in generation t. Then

$$\begin{aligned} f_{t+1} &= \left(1 - \frac{1}{2N_t}\right) f_t + \frac{1}{2N_t} \\ 1 - f_{t+1} &= \left(1 - \frac{1}{2N_t}\right)(1 - f_t) \\ 1 - f_{t+K} &= \left(\prod_{i=1}^{K} \left(1 - \frac{1}{2N_{t+i}}\right)\right)(1 - f_t) \quad . \end{aligned}$$

Now if the population size were constant

$$\left(\prod_{i=1}^{K} \left(1 - \frac{1}{2N_{t+i}}\right)\right) = \left(1 - \frac{1}{2N_e^{(f)}}\right)^K \quad .$$

Dealing with products and powers is inconvenient, but if we take the logarithm of both sides of the equation we get something simpler:[23]

$$\sum_{i=1}^{K} \log\left(1 - \frac{1}{2N_{t+i}}\right) = K \log\left(1 - \frac{1}{2N_e^{(f)}}\right) \quad .$$

It's a well-known fact[24] that $\log(1-x) \approx -x$ when x is small. So if we assume that N_e and all of the N_t are large,[25] then

$$\begin{aligned} K\left(-\frac{1}{2N_e^{(f)}}\right) &= \sum_{i=1}^{K} -\frac{1}{2N_{t+i}} \\ \frac{K}{N_e^{(f)}} &= \sum_{i=1}^{K} \frac{1}{N_{t+i}} \\ N_e^{(f)} &= \left(\left(\frac{1}{K}\right)\sum_{i=1}^{K} \frac{1}{N_{t+i}}\right)^{-1} \end{aligned}$$

The quantity on the right side of that last equation is a well-known quantity. It's the *harmonic mean* of the N_t. It's another well-known fact[26] that the harmonic mean of a series of numbers is always less than its arithmetic mean. This means that genetic drift may play a much more imporant role than we might have imagined, since the effective size of a population will be more influenced by times when it is small than by times when it is large.

Consider, for example, a population in which N_1 through N_9 are 1000, and N_{10} is 10.

$$\begin{aligned} N_e &= \left(\left(\frac{1}{10}\right)\left(9\left(\frac{1}{1000}\right) + \left(\frac{1}{10}\right)\right)\right)^{-1} \\ &\approx 92 \end{aligned}$$

versus an arithmetic average of 901. So the population will behave with respect to the inbreeding associated with drift like a population a tenth of its arithmetic average size.

[23] OK. I know it doesn't look any simpler, but trust me it is. We can work with this one. The other one we can only stare at.

[24] Well known to some of us at least.

[25] So that there reciprocals are small

[26] Are we ever going to run out of well-known facts? Probably not.

Variation in offspring number

I'm just going to give you this formula. I'm not going to derive it for you.[27]

$$N_e^{(f)} = \frac{2N-1}{1+\frac{V_k}{2}} \quad ,$$

where V_k is the variance in number of offspring among individuals in the population. Remember I told you that the number of gametes any individual has represented in the next generation is a binomial random variable in an ideal population? Well, if the population size isn't changing, that means that $V_k = 2(1-1/N)$ in an ideal population.[28] A little algebra should convince you that in this case $N_e^{(f)} = N$. It can also be shown (with more algebra) that

- $N_e^{(f)} < N$ if $V_k > 2(1-1/N)$ and
- $N_e^{(f)} > N$ if $V_k < 2(1-1/N)$.

That last fact is pretty remarkable. Conservation biologists try to take advantage of it to decrease the loss of genetic variation in small populations, especially those that are captive bred. If you can reduce the variance in reproductive success, you can substantially increase the effective size of the population. In fact, if you could reduce V_k to zero, then

$$N_e^{(f)} = 2N-1 \quad .$$

The effective size of the population would then be almost twice its actual size.

References

[1] J F Crow and M Kimura. *An Introduction to Population Genetics Theory.* Burgess Publishing Company, Minneapolis, Minn., 1970.

[2] R Fu, A E Gelfand, and K E Holsinger. Exact moment calculations for genetic models with migration, mutation, and drift. *Theoretical Population Biology*, 63:231–243, 2003.

[27] The details are in [1], if you're interested.

[28] The calculation is really easy, and I'd be happy to show it to you if you're interested.

Creative Commons License

These notes are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.