BIOB480/BIOE548 notes 9/26/2024

Introduction

- More on next week: Actually, I'll be around on Tuesday after all. We'll cover the next topic (small populations); Thursday I'll still be gone, and we'll do the take-home quiz as planned.
- Will have HW3 graded and quiz corrections entered by EOD Friday.
- Questions on HW4?

Cheviron and Brumfield Discussion

See `10_slides.pdf` for questions.

Genetic Drift

The next assumption of HWP we will violate is that of infinite population sizes. When populations are sufficient large in size, the allele frequency dynamics are more or less deterministically driven by mutation, migration, and selection. With finite populations (and particularly with small populations), chance events begin to take on greater importance. This is of particular interest to us because conservation genetics is, in essence, the population genetics of small populations of conservation concern. In Montana, for instance, there only ~350 Black Footed Ferrets remaining, only ~500 Kootenai sturgeon, and only ~320 whooping cranes. The International Union for the Conservation of Nature (IUCN) considers a species critically endangered if there are fewer than 50 individuals, endangered if there are fewer than 250, and vulnerable if there are fewer than 1000.

The primary mechanism for the shift away from deterministic dyanmics is sampling error of gametes. Consider two crosses between sexually reproducing diploid individuals at a single triallelic locus: an A_1A_1 x A_1A_2 cross, and an A_1A_3 x A_1A_2 cross. In this parental generation, allele frequencies are $p = 0.5$, $q = 0.375$, and $r = 0.125$. The first pair then has three offspring, formed by each parent randomly contributing one of their two alleles. The genotypes of these offspring are A_1A_1, A_1A_2, and A_1A_2. The second pair has a single child, with genotype A_1A_1. While this new generation has the same population size and total number of alleles or chromosomes, allele frequencies have changed to $p = 0.75$, $q = 0$, and $r = 0$.

You will notice that since $r = 0$, the allele has disappeared from the population. We describe this as its *loss*: the alternative scenario, in which an allele reaches a frequency of 1.0, is termed *fixation*. To develop intuition about the relationship between population size and the probability of fixation or loss of an allele, using a simulation tool (like CJ Battey's driftR) is helpful. More formally, we can we model the expected distribution of allele frequencies for a given number of gametes using the *binomial distribution*.

Let's start by revisiting Hardy-Weinberg proportions. As usual, we have two alleles (A_1 and A_2) at frequencies p and q. What is the probability that a single individual ($N = 1$) will have a particular genotype? We can determine this by the binomial expansion of $(p + q)^{2N} = (p + q)^2$. The reason is that we have two chances (one allele from both mom and dad, i.e. the exponent) at one of two possible outcomes (receiving either allele A_1, which will occur with a probability proportional to its frequency, p, or A_2, which will occur with a probability proportional to q):

$$(p + q)^2 = (p + q)(p + q) = p^2 + pq + qp + q^2$$

We see our for outcomes: p^2, pq, qp, and q^2. (Typically, $pq + qp$ are lumped into the more familiar $2pq$.) For a single individual, the probability (which you can think of as "expected frequency if there were only one outcome") of having each genotype is therefore $\frac{1}{4} = 0.25$. But what if we have 2 individuals instead? Our binomial expansion will then be raised to a power of $2N = 4$:

$$(p+q)^4 = (p+q)(p+q)(p+q)(p+q) = p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4$$

You can see how this quickly gets complicated. The probability of getting 4 copies of A_1 (out of 4 total alleles in 2 individuals) is p^4, the probability of getting 3 ps and 1 q is $4p^3q$, etc. But how can we generalize this for larger population sizes without the prohibitively tedious process of manually solving the binomial expansion? The formula for the binomial distribution provides us with a shortcut:

$$\binom{2N}{K} p^K q^{2N-K}$$

This formula gives you the probability of there being exactly K copies of an allele with frequency p given $2N$ gametes. Here, $\binom{2N}{K}$ is read "2N choose K" and is equivalent to $\frac{2N!}{K!(2N-K)!}$. q is the frequency of the alternate allele (the binomial expansion only works for diallelic loci). We can return to our example above to demonstrate its use. What, then is the probability of getting 3 copies of allele A_1 given $2N = 4$ and allele frequencies of $p = 0.5$ and $q = 0.5$?

$$\binom{4}{3} 0.5^3 0.5^{4-3} = \frac{4!}{3!(4-3)!} 0.5^3 0.5^1 = \frac{4*3*2*1}{3*2*1*(1)} 0.125 * 0.5 = 0.25$$

Importantly, this is exactly equal to the term in the binomial expansion demonstrated above: $4p^3q = 4 * 0.5^3 * 0.5 = 0.25$.

The variance of the distribution is $\sigma = \frac{p_0 q_0}{2N}$, or the initial allele frequencies over the number of alleles in the population. It is maximized when $p_0 = q_0$ and $2N = 2$, which tells us genetic drift (random change in alelle frequencies due to sampling error) is strongest in small populations with relatively equal frequencies at diallelic alleles. You can build intuition about the binomial distribution by playing around with frequencies ("probability of success") and number of alleles ("number of Bernoulli trials") using this app: https://istats.shinyapps.io/BinomialDist/.

A simple extension of the binomial distribution is that the probability of fixation of an allele is given by its frequency to the $2N$ power, i.e. p^{2N}. Conversely, the probability of its loss is its complement to the $2N$ power, ie. $(1-p)^{2N}$.