An Example from Population Genetics: The Wright-Fisher Model

Today we consider a stochastic process used to describe the way genes get transmitted from one generation to the next in an ideal population called a Wright-Fisher population. Our agenda is:

1. describe the "physical model" of a W-F population (brief)
2. translate the physical model into mathematical terms
3. compare this process to Gambler's Ruin to get a feel for some of its new features
4. answer several questions about the process—there are many things that one might want to know about this process. Some of those things are quite difficult to determine. We will stick to some elementary results which will lead us to jog our memories on conditional expectation and variance, and to solve a very simple inhomogeneous recurrence equation.

Physical Model

Physically, we consider a population of constant size N diploid organisms (this means there are $2N$ genes) each generation where generations are indexed by $n = 0, 1, \ldots$. Each organism lives only one generation. Everyone dies right after the offspring are made. The mating scheme and offspring survival scheme are such that one can think of each individual in the next generation receiving two genes, each one selected randomly and with replacement from the genes present among the parents. In fact, with this sort of system for the questions we wish to answer today, one really needn't even consider the individuals that the genes get placed into. It suffices to think of the population consisting of $2N$ genes some number x, $(0 \leq x \leq 2N)$ of which are type A genes. The next generation, then, will be $2N$ genes, sampled with replacement from the $2N$ in the previous generation.

The Process, Mathematically

- $2N$ genes
- at time $n = 0$, x of these genes are type A, $(0 \leq x \leq 2N)$
- Y_n is the random variable for the number of A genes at time n, ($y_n =$ realized value)
- $(Y_n | Y_{n-1} = y_{n-1}) \sim \text{Bin}(2N, y_{n-1}/2N)$ (sampling with replacement)
- Hence

$$P(Y_n = y_n | Y_{n-1} = y_{n-1}) = \frac{2N!}{y_n!(2N - y_n)!} \left(\frac{y_{n-1}}{2N}\right)^{y_n} \left(1 - \frac{y_{n-1}}{2N}\right)^{2N - y_n} \tag{1}$$

Comparison to Random Walks You've Studied Earlier

In several ways this is a somewhat more complicated example than either the Gambler's Ruin (GR) or the Prisoner's Escape (PE). Most notably:

- Unlike GR, the transition probabilities depend on the current state

  - Nothing really shocking here—you've seen that in PE

- The Wright-Fisher process may make steps of many different sizes—unlike GR (or simple random walk) where steps were either up one or down one, and also unlike PE where steps were either up one or all the way back to zero. To really appreciate this, let's make some "from-to" tables of the transition probabilities for both W-F and GR. If we write $P_{i,j}$ for $P(Y_n = j|Y_{n-1} = i)$, then our table for the Wright-Fisher process looks like:

$$\begin{array}{c} \nearrow \\ 0 \\ 1 \\ \vdots \\ 2N-1 \\ 2N \end{array} \begin{pmatrix} 0 & 1 & \cdots & 2N \\ 1 & 0 & \cdots & 0 \\ P_{1,0} & P_{1,1} & \cdots & P_{1,2N} \\ \vdots & \vdots & \ddots & \vdots \\ P_{2N-1,0} & P_{2N-1,1} & \cdots & P_{2N-1,2N} \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

  while the table for the Gambler's Ruin on $\{0, 1, \ldots, N\}$ looks like:

$$\begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ N-2 \\ N-1 \\ N \end{array} \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & \cdots & N-2 & N-1 & N \\ 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & q & 0 & p & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & q & 0 & p & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & q & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}$$

- One similarity between GR and W-F is in the absorbing states on the ends. It is certain in both processes that it will end up in one of the endoints (0 or 2N for W-F and 0 or N for GR) as $n \to \infty$.

Describing Quantities re: The Behavior of a Wright-Fisher Model

It would be useful to know several things about the behavior Y_n in the W-F process. I would love to have a quick and easy way of computing $P(Y_n = y_n|Y_0 = x)$, but, alas, there is no way of doing this without summing over all the intermediate steps the process might have taken. We settle instead for finding expressions for (1) π_x, the probability of extinction of a gene given it started with x copies, (2) the expected value $E(Y_n)$ and (3) the variance $\text{Var}(Y_n)$.

Finding π_x:

- Let $\pi_x = \lim_{n\to\infty} P(Y_n = 0|Y_0 = x)$
- "Knee Jerk" option number one $\longrightarrow$ try a first step analysis
  - This will look like: $\pi_x = 1 \cdot P_{x,0} + \pi_1 P_{x,1} + \pi_2 P_{x,2} + \pi_3 P_{x,3} + \pi_3 P_{x,3} + \cdots + \pi_{2N-1} P_{x,2N-1} + 0 \cdot P_{x,2N}$

- A long and wholly unsavory recurrence equation. This is NOT how we want to go about this!
- We defer this until we know $E(Y_n)$

Finding $E(Y_n)$ (very easy!)

- Use the expectation of a conditional expectation:

$$E(Y_n) = E[E(Y_n|Y_{n-1})] = E(Y_{n-1}) = E(Y_{n-2}) = \cdots = E(Y_0) = x$$

Finding π_x revisited (very easy!):

- Now, $\lim_{n\to\infty} E(Y_n) = x$, so

$$x = 0 \cdot \pi_x + 2N \cdot (1 - \pi_x) \Longrightarrow \pi_x = \frac{2N - x}{2N}$$

- This is the same as alternative (iii) when $p = q$ in GR process

Finding a tidy expression for $\mathrm{Var}(Y_n)$

The strategy here is to find $\mathrm{Var}(Y_n)$ in terms of $\mathrm{Var}(Y_{n-1})$ and then solve the resulting, simple inhomogeneous recurrence equation. Note that this is going to depend on x, the initial number of type A genes.

- Recall binomial variance: if $X \sim \mathrm{Bin}(N, p)$ then $\mathrm{Var}(X) = Np(1-p)$
- Employ the useful fact that $\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}[E(Y|X)]$:

$$\begin{aligned}
\mathrm{Var}(Y_n) &= E[\mathrm{Var}(Y_n|Y_{n-1})] + \mathrm{Var}[E(Y_n|Y_{n-1})] \\
&= E\left[2N\tfrac{Y_{n-1}}{2N}(1 - \tfrac{Y_{n-1}}{2N})\right] + \mathrm{Var}(Y_{n-1}) \\
&= E\left[Y_{n-1}(1 - \tfrac{Y_{n-1}}{2N})\right] + \mathrm{Var}(Y_{n-1}) \\
&= E(Y_{n-1}) - \tfrac{1}{2N}E(Y_{n-1}^2) + \mathrm{Var}(Y_{n-1}) \\
&= x - \tfrac{1}{2N}[\mathrm{Var}(Y_{n-1}) + (EY_{n-1})^2] + \mathrm{Var}(Y_{n-1}) \\
&= x - \tfrac{1}{2N}[\mathrm{Var}(Y_{n-1}) + x^2] + \mathrm{Var}(Y_{n-1}) \\
\mathrm{Var}(Y_n) &= (1 - \tfrac{1}{2N})\mathrm{Var}(Y_{n-1}) + x\left(1 - \tfrac{x}{2N}\right) \qquad (2)
\end{aligned}$$

- Equation 2 gives us our desired inhomogeneous recurrence equation

- It's easier to solve the recurrence for the quantity $\text{Var}(Y_n) - 2Nx(1 - \frac{x}{2N})$:

$$\text{Var}(Y_n) - 2Nx(1 - \tfrac{x}{2N}) = (1 - \tfrac{1}{2N})\text{Var}(Y_{n-1}) + x(1 - \tfrac{x}{2N}) - 2Nx(1 - \tfrac{x}{2N})$$

  which gives, after some rearrangement

$$\text{Var}(Y_n) - 2Nx(1 - \tfrac{x}{2N}) = (1 - \tfrac{1}{2N})[\text{Var}(Y_{n-1}) - 2Nx(1 - \tfrac{x}{2N})]$$

  and with the boundary condition $\text{Var}(Y_0) = 0$ we have finally:

$$\text{Var}(Y_n) = 2Nx\left(1 - \frac{x}{2N}\right)\left[1 - \left(1 - \frac{1}{2N}\right)^n\right]$$

Check Results On $n = 1$ and $n \to \infty$ and interpret