The Wright−Fisher model of Genetic Drift and the Hardy−Weinberg principle explained

F. Borgonovo, DEIB, Politecnico di Milano.

1 Introduction

This note is addressed to students and other inquisitive people that want to approach the field of Population Genetics browsing the Web to learn basic ideas and results. Unfortunately, the many Web pages that deal with the Hardy−Weinberg principle, the Wright−Fisher model and Genetic Drift, are often plagued by misinterpretations, bad language/exposition and lack of accuracy that can ingenerate confusion. In fact, Population Genetics has developed, for historical reasons, a language and denominations that sometimes are in conflict with the terms used in probability theory, whom the cited topics use in large measure. Here we address such topics detailing the underlying framework with the correct probabilistic assumptions and terms, show how the main results are derived, provide the correct interpretations and underline the critical misinterpretation points.

We give for granted the knowledge of the basic probability theory and the Mendel's laws. We do not provide examples or graphs since many of them can be retrieved in the Internet. We do not give references about the topic investigated and the the probabilistic tools we use, i.e., Markov Chain, since they are common teaching material whose references, again, can easily be retrieved in the Internet.

2 Assumptions

We refer to the example usually dealt with in Population Genetics, where a population of individuals gives rise to subsequent generations all with the same number individuals N. Each element of the population has a *genotype* composed of two alleles of a gene that can be of two different types, say A and a, (e.g. dominant and recessive). The way in which each individual at generation

$n+1$ inherits its two alleles from the pool of alleles at generation n depends on the inheritance laws that are assumed. We postulate two such laws:

Law 1. *Each child allele at generation $n+1$ is selected at random within the pool of alleles at generation n.*

Law 2. *Each child genotype is inherited according to the Mendel's laws by randomly selecting the parent's geneotypes.*

No other effect influences the allele next generation outcome.

According to Law 1, if the genes A and a are in numbers r and $2N-r$, the offspring probilities are respectively

$$p_A = \frac{r}{2N}, \qquad p_a = \frac{2N-r}{2N}, \qquad p_A + p_a = 1. \tag{1}$$

Now assume that genotypes are distributed according to the numbers $u,\ v,\ z,\ u+v+z=N$, with respect to genotypes (AA), (Aa), and (aa) respectively. Thus the genotypes are sected with probabilities

$$p_{AA} = \frac{u}{N}, \qquad p_{Aa} = \frac{v}{N}, \qquad p_{aa} = \frac{z}{N}, \qquad p_{AA} + p_{Aa} + p_{aa} = 1. \tag{2}$$

The probability that an inherited allele is of type A, and a, are given by

$$p_A = p_{AA} + p_{Aa}\frac{1}{2}, \qquad p_a = p_{aa} + p_{Aa}\frac{1}{2}, \qquad p_A + p_a = 1, \tag{3}$$

which, of course, are exactly the fraction of the corresponding number of alleles in the population:

$$p_A = \frac{2u+v}{2N}, \qquad p_a = \frac{2z+v}{2N}, \qquad p_A + p_a = 1. \tag{4}$$

Now, by Law 2 (we assume for granted Mendel's laws) the probability that the inherited genotype is an AA type can be evaluated as

$$p'_{AA} = p^2_{AA} + p_{AA}p_{Aa} + p^2_{Aa}\frac{1}{4} = \left(p_{AA} + p_{Aa}\frac{1}{2}\right)^2 = p^2_A, \tag{5}$$

where the last equality comes from (3). In the same way we can prove that

$$p'_{Aa} = 2p_A p_a, \qquad p'_{aa} = p^2_a. \tag{6}$$

Eqs. (5) and (6) prove the following

Property 1. *The distribution of alleles in a genotype inherited with Law 2 is the same we get by selecting at random and independently the alleles in the genotypes with Law 1.*

Since by Property 1 the statistics of genotypes can be derived from the distribution of alleles by (5) and (6), to study the evolution of genotypes generation by generation we can simply refer to the evolution of alleles, that depend on the simpler Law 1. This law, and the assumption that generations do not overlap, yield the model known in the literature as the *Wright−Fisher model*, which is analized next.

3 Analysis

3.1 The probability distribution

According to the Wright−Fisher model, we study the random variable $N_A(n)$, defined as the number of alleles of type A at generation n, being the number of type a allele given by $2N - N_A(n)$. The random selection of parents at different generations N_A changes from one generation to the next, giving rise to a sequence of random variables $N_A(0)$, $N_A(1)$, $N_A(2)$, $\ldots N_A(n)\ldots$, that in probability theory is called a *stochastic process*.

If we are given many populations, each population with the same number $2N$ of alleles starting at generation 0 with the same number $N_A(0) = r$ of type A alleles, and using the same inheritance laws, we turn out with many sequences, one for each population:

$$r,\ \ N'_A(1),\ N'_A(2),\ \ldots\ N'_A(n),\ldots$$

$$r,\ \ N''_A(1),\ N''_A(2),\ \ldots\ N''_A(n),\ldots$$

$$r,\ \ N'''_A(1),\ N'''_A(2),\ \ldots\ N'''_A(n),\ldots$$

Not only the values at different generations, $n > 0$, of each sequence, are usually different, but they are different also with respect to populations because, again, the random selection mechanism operates randomly and independently in each population. Genetists call the sequences (1) *samples*, whereas mathematician call them *realizations* or *sample paths* of the process.

Actually the process $N_A(n)$ under consideration is a very specific one, known by mathematicians as Markov Chain. This chain is probabilistically described at each generation n by the row vector $\mathbf{\Pi}(n)$ composed of the probability distribution of the number N_A, i.e., the vector's $2N+1$ elements

are:

$$Pr(N_A(n) = 0),\ Pr(N_A(n) = 1),\ Pr(N_A(n) = 2), \ldots$$

$$\ldots Pr(N_A(n) = 2N - 1),\ Pr(N_A(n) = 2N). \tag{7}$$

where $Pr(N_A(n) = k)$ is the probability that the number of type A allele at generation n is equal to k.

If we start at generation 0 with a given number $N_a(0) = r$ of type A alleles, generation 1 is attained, by Law 1, as if flipping coins with the A outcome having probability $p_1 = p_A = r/2N$. Therefore, we can say

Property 2. *The distribution of $N_A(1)$, at generation 1, is given by the Binomial distribution of parameters r and $p_1 = r/2N$:*

$$Pr(N_A(1) = k) = \binom{2N}{k} p_1^k (1 - p_1)^{n-k}, \qquad 0 \leq k \leq 2N. \tag{8}$$

Things become a little more complex for subsequent generations. Shortly, if in a sample path at generation n we observe $N_A(n) = i$, then again the conditional distribution at generation $n + 1$ is given by the Binomial distribution of parameters i and $p_n = i/2N$. In order to get the distribution, over all sample paths, we must average all the conditional distributions (vectors) at $n + 1$ with the distribution at time n. This means building the matrix $\mathbf{Q}$ (the transition matrix) whose rows are composed by the conditional distributions for all i; then distribution $\mathbf{\Pi}(n + 1)$ is attained as the following matrix product

$$\mathbf{\Pi}(n + 1) = \mathbf{\Pi}(n)\mathbf{Q}. \tag{9}$$

Therefore we have:

Property 3. *The probability distribution at any time n is attained by recursively applying* (9) *starting from row vector:*

$$\mathbf{\Pi}(0) = [0\ 0\ 0 \ldots 0\ 1\ 0 \ldots 0\ 0], \tag{10}$$

where the only 1 *is in position r.*

In practice the evaluation complexity of (9) is such that we can numerically derive the distribution, generation by generation, only if N is not too large, up to one hundred or so. The theory also assures that the Markov Chain $N_A(n)$ reaches an n^* (that changes from sequence to sequence) after which the chain will reach *absorbing states*, i.e., either $N_A(n) = 2N$, where all alleles will be

of the A type and no further change is possible, or $N_A(n) = 0$, where all alleles will be of the a type and no further change is possible (genetists call this *fixation* and *loss* of A respectively, and the reverse for allele a). This result can be summarized into

Property 4. *The number of alleles changes with generations, and ends up either with $N_A = 2N$ or $N_A = 0$.*

This means that for all $n > n^*$ the distribution vector in (7) is always the same and equal to:

$$1-\alpha, \quad 0, \quad 0, \quad \dots \quad 0, \quad \alpha \tag{11}$$

where $\alpha = Pr(N_A(n) = 2N)$ represents the probability that type A allele becomes fixed. Of course the actual value of α depends on $N_A(0)$, the number of type A alleles at generation 0, and increases, from 0 to 1 as $N_A(0)$ increses from 0 to $2N$. The value of α is determined as we explain later in Section 3.3.

An interesting case in genetics is when $N_A(0) = 1$, modeling the case of *mutation*, where at generation 0 a new allele appears among a population of $2N$ equal alleles. The probability that the mutated allele is lost after one generation is the probability that none the $2N$ individuals of the following generation selects the mutated allele, i.e., the probability in (8) with $k = 0$ and $p_1 = 1/2N$. Then we have

$$Pr(N_A(1) = 0) = \left(1 - \frac{1}{2N}\right)^{2N} \longrightarrow e^{-1} \approx 0.368,$$

where the limit on the left is attained for $N \to \infty$. Therefore, for large populations, the probability that a mutation is loss immediately after being generated is almost 37%.

Quite often instead of referring to N_A, the literature refers to $f_A = N_A/2N$, which is called the *frequency* of allele A, and by genetists the *sample frequency* or the *observed frequency*. Dividing by the constant $2N$ does change the units but does not change the dynamic of the process and, therefore, the values of the probability distribution of the frequency f_A is simply derived as

$$Pr(f_A = i/2N) = Pr(N_A = i),$$

so that results about frequencies are immediately derived from the results about N_A. In term of frequencies Property 4 becomes:

Property 5. *The frequency of alleles changes with generations, and ends up either with $f_A = 1$ or $f_A = 0$.*

3.2 The average of the distribution

Property 3 shows that the distribution at generation $n = 1$ can be easily derived as the Binomial distribution ((8)). The average (expectation) of the Binomial distribution is

$$\mathsf{E}\left[N_A(1)\right] = 2Np_1 = r,$$

that is, $N_A(1)$ changes over different sample paths, but its average is still equal to r, i.e. the value of $N_A(0)$. This is a direct consequence of the inheritance law given a precise number r of type A alleles; how about next generations where $N_A(n)$ no longer corresponds to a precise number being a random variable? The answer is in the following

Property 6. *The average number of alleles does not change with generations.*

Therefore we have

$$\mathsf{E}\left[N_A(n)\right] = r, \qquad \forall n, \tag{12}$$

even though $N_A(n)$, for $n > 1$, is no longer binomial. The proof is based on the conditional expectation, symbolized by the formula

$$\mathsf{E}\left[N_A(n+1)\right] = \mathsf{E}\left[\mathsf{E}\left[N_A(n+1)|N_A(n)\right]\right]. \tag{13}$$

In deriving (9) we saw that, once we know the outcome of generation n, say $N_A(n) = x$, the outcome at $n+1$ is still a binomial, whose average is

$$\mathsf{E}\left[N_A(n+1)|N_A(n)\right] = 2Np_{n+1},$$

and being $p_{n+1} = N_A(n)/2N$ we get

$$\mathsf{E}\left[N_A(n+1)|N_A(n)\right] = N_A(n).$$

Finally (13) yields

$$\mathsf{E}\left[N_A(n+1)\right] = \mathsf{E}\left[N_A(n)\right]. \qquad \forall n. \tag{14}$$

Result (12) comes by recursively applying the above starting from $n = 1$.

Dealing with frequencies, which means dividing by $2N$, we have

$$\mathsf{E}\left[f_A(n)\right] = \frac{r}{2N} = p_1, \qquad \forall n, \tag{15}$$

where the last passage comes from (1), since $p_1 = p_A$. Therefore we can state

Property 7. *The average frequency of alleles does not change with generations.*

3.3 The allele fixation

After the fixation of A, or a, has occurred the expectation is

$$\mathsf{E}\left[N_A(n)\right] = \alpha 2N \qquad n > n^*.$$

and, by Property 6, and (12), it equals r, so that we can derive α as

$$\alpha = r/2N = p_1,$$

and we have

Property 8. *The fixation probability of the allele equals the original frequency of the allele itself.*

As an example, if at $n = 0$ in a population of size $N = 1000$ we have $N_A = 20$, then the probability that A becomes fixed is 10%, whereas 90% is the probability that allele a becomes fixed.

Markov Chain theory also teaches how to evaluate the average times T it takes to absorption (either fixation or loss of the allele), and the conditional average time to fixation (absorption) T_A given that fixation occurs. Unfortunately no close form exists for such times; however, the following approximated formulas have been proved to provide good results:

$$T = -4N\left(p_1 \ln p_1 + (1-p_1)\ln(1-p_1)\right), \tag{16}$$

$$T_A = -4N\frac{(1-p_1)\ln(1-p_1)}{p}. \tag{17}$$

The value of T is (obviously) maximum for $p_1 = 0.5$, where we have $T \approx 2.8N$, and is zero for $p_1 = 0$. When we have a single type A allele, as it happens with mutations, the probability it becomimg fixed is $p_1 = 1/2N$ (Property 8), and (17) correspondingly provides ($p_1 \to 0$), $T_A \approx 4N$.

3.4 The variance of the distribution

We have seen that the average number of alleles does not change with generations, but the distribution does, starting from (10) and ending in (11). A measure of this variation is the variance of random variable $N_A(n)$:

$$\mathrm{VAR}\left[N_A(n)\right] = \mathsf{E}\left[\left(N_A(n) - \mathsf{E}\left[N_A(n)\right]\right)^2\right].$$

Using (10) and (11) we get

$$\mathrm{VAR}\left[N_A(0)\right] = 0, \qquad \mathrm{VAR}\left[N_A(n)\right] = 4p_1N^2(1-p_1),\ n > n^*.$$

and for the frequency:

$$\mathrm{VAR}\left[f_A(0)\right] = 0, \qquad \mathrm{VAR}\left[f_A(n)\right] = p_1(1-p_1),\ n > n^*.$$

For intermediate values of n the variance can be approximated as

$$\mathrm{VAR}\left[f_A(n)\right] \approx p_1(1-p_1)\left(1-e^{-n/2N}\right). \tag{18}$$

3.5 Large numbers

The main law that makes Probability Theory useful is the law of large numbers. This law states that the frequency N_X/N of an observation X in N repetitions of an experiment tends, as N increases, to the probability p_X that such an observation occurs in a single experiment. We write this as:

$$lim_{N\to\infty}\frac{N_X}{N} = p_X.$$

Actually the meaning of the *lim* operation above is complex to explain. However, what is relevant here is that, for large N, we can assume:

$$\frac{N_X}{N} \approx p_X.$$

In our case, N is the population size, actually $2N$ if we refer to alleles. If in a generation we observe N_A occurrences of the type A allele under a given inheritance law then, if N is very large, the frequency $N_A/2N$ practically coincides with p_A, the probability of occurrence of allele A in a single experiment on the inheritance law. Then, if we consider the multiple paths of our genetic process, all the frequencies over all paths tends to the same (frequency) p_A, which also coincides with the average frequency. This is also confirmed by the variance (18) going to zero when N goes to infinity. Property 7 then becomes

Property 9. *In an infinite population the frequency of alleles does not change with generations.*

4 The genetic drift

The change of the frequency with generations is what genetists call *genetic drift*. In Property 5 we have stressed the fact that the frequency of alleles *always* changes with generations. This is to warn against statements often found in Population Genetics affirming that for large populations

the frequency of alleles does not change, which is not mathematically correct except with an infinite population size (Property 9), which does not exists. However, the difference in the variance (18) between two generations is

$$\begin{aligned}&\mathrm{VAR}\left[f_A(n+1)\right]-\mathrm{VAR}\left[f_A(n)\right]\approx p_1(1-p_1)\left(e^{-n/2N}-e^{-(n+1)/2N}\right)\\&=p_1(1-p_1)e^{-n/2N}\left(1-e^{-1/2N}\right)\approx p_1(1-p_1)e^{-n/2N}\frac{1}{2N}.\end{aligned}\tag{19}$$

where the last approximation holds for large population size N, where $1/2N << 1$. Therefore we understand that for $N > 1000$ this change can be neglected, and for still larger N the change can be neglected over several generations.

In the case of a new mutated allele, as seen in Section 3, the probability it becomes fixed is $1/N$, very small indeed for large N, which shows that some other mechanisms must intervene to help the allele fixation, and increase *genetic variation*, as observed in nature. The most accepted effects that help fixation are the *bottleneck effect* and the *founder effect*. In both cases the population happens to be restricted to a small number N, which makes fixation more probable. When the population size increases, also because of migrations, *natural selection* causes an increases in the frequency of an allele that makes individuals more *fit* to survivability.

5 The Hardy−Weinberg principle

In the *Wright−Fisher model* with an infinite population the frequencies of alleles do not change from generation to generation (Property 9). Hardy was asked what about the frequency of genotypes.

Given the genotype population probability distribution as in (2), under the Mendel's laws and the other conditions of the the *Wright−Fisher model*, the genotype probability distribution of a new generation individual is given by

$$p'_{AA}=p_{AA}^2+p_{AA}p_{Aa}+p_{Aa}^2\frac{1}{4}=\left(p_{AA}+p_{Aa}\frac{1}{2}\right)^2,\tag{20}$$

$$p'_{Aa}=2\left(p_{AA}+p_{Aa}\frac{1}{2}\right)\left(p_{aa}+p_{Aa}\frac{1}{2}\right),\tag{21}$$

$$p'_{aa}=p_{aa}^2+p_{aa}p_{Aa}+p_{Aa}^2\frac{1}{4}=\left(p_{aa}+p_{Aa}\frac{1}{2}\right)^2,\tag{22}$$

being (20) equal to (5).

Because it is $p_{AA} + p_{Aa} + p_{aa} = 1$, and due to the specific form of (20)-(22), in order to get the same distribution, i.e. $p' = p$, we need only to impose the equality to one of them. If we impose $p'_{Aa} = p_{Aa}$ using the above we have

$$p_{Aa}\frac{1}{2} = \left(p_{AA} + p_{Aa}\frac{1}{2}\right)\left(p_{aa} + p_{Aa}\frac{1}{2}\right) = p_{Aa}\frac{1}{2}(p_{AA} + p_{aa}) + p_{AA}p_{aa} + p_{Aa}^2\frac{1}{4}$$

$$= p_{Aa}\frac{1}{2}(1 - p_{Aa}) + p_{AA}p_{aa} + p_{Aa}^2\frac{1}{4}$$

which provides

$$\frac{1}{4}p_{Aa}^2 = p_{AA}p_{aa}. \tag{23}$$

The (23) is known as the Hardy−Weinberg's condition.

We note that we must explicitly select the distribution u, v and z at generation 0 if we want condition (23) to be met, i.e., if we want the distributions at generation 0 and 1 be the same. However, whatever the distribution u, v and z at generation 0, Property 1 assures that the distribution at generation 1 obey the Hardy−Weinberg's condition (23). In fact, for such a population, owing to (5) and (6), the two terms of (23) become

$$\frac{1}{4}p_{Aa}^2 = p_A^2 p_a^2$$

$$p_{AA}p_{aa} = p_A^2 p_a^2.$$

This assures that the distribution at generation 2 and all the subsequent, are equal to the distribution at generation 1, whatever the distribution at generation 0. This can be resumed into

Property 10. *When the three genotypes are derived by randomly and independently selcting the alleles of each genotype according to Property 1, the Hardy− Weinberg's condition* (23) *is satisfied.*

This means that Properties 7 and 9 also holds for genotype frequencies:

Property 11. *When generations of an infinite population are derived according to Law 2, the genotype frequencies $N_{AA}(n)/N$, $N_{Aa}(n)/N$ and $N_{aa}(n)/N$, do not change with all $n \geq 1$.*

Property 12. *When generations of a finite population are derived according to Law 2, the genotype average frequencies $\mathsf{E}\left[N_{AA}(n)/N\right]$, $\mathsf{E}\left[N_{Aa}(n)/N\right]$ and $\mathsf{E}\left[N_{aa}(n)/N\right]$, do not change with all $n \geq 1$.*

As often stated in the literature, the Hardy−Weinberg principle is explained as guaranteeing Proposition 11; actually, this is rather restrictive with respect to the principle implications. In

fact it implies Proposition 12, which does not need the infinite population assumptions. The real meaning of the principle lies in its reverse, namely the fact that,

Property 13. *If the condition of Hardy− Weinberg is not satisfied, then the Wright−Fisher model must be rejected.*

Given that the Mendel's laws constitute a well established fact, a failure of the Hardy−Weinberg condition on real populations indicates that either mating is not casual or other non-random selection ways, such as natural selection, intervene changing the genotype frequency in the population.

Published January 2021.