EXAMINATION OF POPULATION GENETICS AND HARDY-WEINBERG EQUILIBRIUM PRINCIPLES

CANDIS KAPUSCINSKI
DR. VOCHITA MIHAI

Abstract. Population genetics studies the genetic composition of a population. Population genetics models investigate the occurrence of traits within a population in the past and present in attempt to identify normal trends, make predictions about future genetic distributions, and understand how genetic variability within a population inevitably leads to evolutionary change. In this article we will use difference equations and assume the principles of Hardy-Weinberg Law are upheld in order to prove genetic variation does not change from generation to generation. Further, we will determine when equilibrium exists within the given population. We will apply these difference equations to show that selection within a population leads to an increase in the mean fitness of a population over time.

1. BASIC CONCEPTS AND DEFINTIONS

The father of genetics, Gregor Mendel (1866), used basic mathematics to calculate probabilities of traits in future generations. His experiments with pea plants showed that offspring inherit half of their genetic material from one parent, and half from the other parent. For example, humans receive one set of 23 chromosomes from the father and one set from their mother, for a total of 46 chromosomes. Genes, which Mendel discovered as the instructions for passing on hereditary information from generation to generation in all organism are located on chromosomes at specific points, called loci. Genes at various loci code for certain traits (i.e. eye color, skin tone, hair type, etc.) and often times, multiple genes work together to influence a single trait. Variations of genes are referred to as alleles. For example, if hair color is the gene to be expressed, allelic possibilities include: brown, blonde, black, or red hair.

Date: May 26, 2015.

For our purposes we are only referring to human cases in which there are only two allelic possibilities (G or g) for a gene. This means we are dealing with a single gene, at a single locus, with only two possibilities. Because offspring obtain one allele from each parent, the offsprings allelic makeup will be one of three possibilities (genotypes): GG, Gg, or gg. When the offspring inherits two of the same alleles, GG or gg, the combination is referred to as homozygous. The combination is referred to as heterozygous when the offspring inherits one of each allele type.

An important concept to understand in genetics is dominance. Dominance refers to the way two alleles interact with one another. Alleles can be either dominant or recessive. In the case of the allele possibilities G or g, we can refer to G as the dominant allele and g as the recessive allele. Complete dominance occurs if one allele in the heterozygous genotype (Gg) completely masks the effect of the other. In this case, the physical expression, or phenotype, will appear identical to that of Gg. Thus, when complete dominance occurs with two allele possibilities, there are two phenotype possibilities for the three different genotypes. In some cases, such as sickle cell disease, codominance occurs. Codominance can be seen in the heterozygous case, Gg. In this case, two different alleles for a trait are both expressed, neither allele is dominant or recessive. When alleles interact codominantly, there are three unique phenotypes for all three genotypes.

2. MATHEMATICAL MODELS

For better understanding of population distributions, mathematical models have been developed and improved. Population genetics and mathematics first mixed in the mid-19th century when Gregor Mendel used elementary mathematics to calculate expected allele frequencies in future generations of pea plants. Later on, Francis Galton and Karl Pearson were able to identify trait distributions within a population and estimate their fluctuation between generations with new statistical procedures. The work of Ronald A. Fisher, J.B.S. Haldane, and Sewall Wright formed the foundation for modern population genetics by showing the theory of evolution by natural selection can be justified by Mendelian principles of genetics through mathematical modeling.

3. HARDY-WEINBERG LAW

The principle we now refer to as the Hardy-Weinberg Law was discovered independently by G.H. Hardy, an English mathematician, and Wilhelm Weinberg, a German physician and geneticist in the early 1900s. They showed that under certain circumstances, genetic variation does not change from generation to generation. More precisely, they showed that genotypic allelic frequencies remain stable as each generation reproduces.

Theorem 3.1. *Hardy-Weinberg Law*
Assume the parent population has two alleles for a particular gene, A and a, and the initial proportion of allele A is $p_0$and the initial proportion of allele a is q_0. In addition, it is necessary to make the following assumptions.

1. *Random mating (takes place without regard to ancestry or genotype)*
2. *All genotypes are equally fit (equal survival probability)*
3. *No mutations*
4. *No variation in the number of progeny from parents of different genotypes*
5. *No immigration or emigration*
6. *Generations are nonoverlapping*

Then, the allelic frequencies in generation t remain constant and the genotypic frequencies remain constant from the second generation to each subsequent generation. That is:

$$p_t = p_0,\ q_t = q_0,\ p_A = p_0^2,\ p_B = 2p_0q_0,\ \textit{and}\ p_C = q_0^2.$$

Proof. The following definitions and equations pertain to the above assumptions. Let N be the total population size.
Since there are two types of alleles per locus, A and a, then $2N$ is the total number

of alleles in the population.

Let p be the proportion of alleles A in population N, this means

$$p = \frac{total\ number\ of\ A\ alleles}{2N}$$

Let q be the proportion of allele a in population N this means

$$q = \frac{total\ number\ of\ a\ alleles}{2N}$$

$$p+q = \frac{total\ number\ of\ A\ alleles\ + total\ number\ of\ a\ alleles}{2N} = \frac{total\ number\ of\ alleles}{2N} = \frac{2N}{2N} = 1$$

thus

$$p + q = 1$$

Further, let A be genotype AA, B be genotype Aa, and C genotype aa.

Let p_A be the proportion of genotype AA, p_B be the proportion of genotype Aa, and p_C the proportion of genotype aa.

If p is the probability of alleles A and q the probability of alleles a then

$$p = \frac{2Np_A + Np_B}{2N} = p_A + \frac{1}{2}p_B$$

and

$$q = 1 - p = \frac{2Np_C + Np_B}{2N} = p_C + \frac{1}{2}p_B$$

When considering all possible mating possibilities, it is determined that there are 6:

(1) AA with probability p_A^2
(2) AB which is the same with BA, with probability $2p_Ap_B$
(3) AC which is the same as CA with probability $2p_Ap_C$
(4) BB with probability p_B^2
(5) BC which is the same as CB, with probability $2p_Bp_C$

(6) CC with probability p_C^2

Let p_A, p_B, and p_C represent genotype probabilities for the 1st generation (after one mating cycle) and p'_A, p'_B, and p'_C represent genotype probabilities for the 2nd generation (the mating of the 1st generations).

Table 1 provides mating possibilities and frequencies as well as the probability of each genotype in the 1st and 2nd generation.

Table 1

Mating types	Mating probability	A_1	B_1	C_1	A_2	B_2	C_2
AA	p_A^2	1	0	0	p_A^2	0	0
AB	$2\mathrm{p}_A p_B$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\mathrm{p}_A p_B$	$\mathrm{p}_A p_B$	0
AC	$2\mathrm{p}_A p_C$	0	1	0	0	$2\mathrm{p}_A p_B$	0
BB	p_B^2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}p_B^2$	$\frac{1}{2}p_B^2$	$\frac{1}{4}p_B^2$
BC	$2\mathrm{p}_B p_C$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\mathrm{p}_B p_C$	$\mathrm{p}_B p_C$
CC	p_C^2	0	0	1	0	0	p_C^2

Where A_1, B_1, C_1 are the first generation offspring probabilities and A_2, B_2, C_2 are second generation offspring probabilities.

Let p'_A, p'_B, p'_C represent probabilities of each genotype in the 2nd generation. Using Table 1, the sum of the probabilities from column 6 gives us:

$$p'_A = p_A^2 + p_A p_B + \frac{p_B^2}{4} = \left(p_A + \frac{p_B}{4}\right)^2 = p^2$$

The sum of probabilities from column 7:

$$p'_B = p_A p_B + 2p_A p_C + \frac{p_B^2}{2} + p_B p_C = p_A\left(p_B + 2p_C\right) + \frac{1}{2}p_B\left(p_B + 2p_C\right) = \left(p_B + 2p_C\right)\left(p_A + \frac{1}{2}p_B\right) = 2pq$$

The sum of the probabilities from column 8:

$$p'_C = \frac{1}{4}p_B^2 + p_B p_C + p_C^2 = (p_C + \frac{1}{2}p_B)^2 = q^2$$

It is possible to find the allelic frequencies of the population in the second generation by using these sums.

$$p' = p'_A + \frac{1}{2}p'_B = p^2 + pq = p(p+q) = p$$

$$q' = p'_C + \frac{1}{2}p_B q^2 + pq = q(q+p) = q$$

Therefore, when taking into account the 6 assumptions of the Hardy-Weinberg law, allelic frequencies remain constant in each generation. Probabilities for p and q at any time, t, are the same at time, t + 1. That is,

$p_t = p_{t+1}$ and $q_t = q_{t+1}$.

The initial allelic probabilities will be the same in each subsequent generation.

$p_t = p_0$ and $q_t = q_0$

□

However, if any of the assumptions of the Hardy-Weinberg Law are violated, this principle does not hold. For example, if the assumption that all genotypes have equal survivability is false, this means that survival rate is dependent on genotype. If this is the case, each genotype has a different level of fitness, denoted by the letter w. Let w_A, w_B, w_C represent the constant survival rates of genotypes A, B, C, respectively. We are assuming $w_B = 1$ and the survival rates w_A and w_C are relative to w_B. Further, the frequency of the dominant allele, A, in generation t will be denoted by p_t and the frequency of the recessive allele, a, will be denoted by q_t.

The mean fitness for generation t is given by the equation:

$$w_t = p_t^2 w_A + 2p_t q_t w_B + q_t^2 w_C \tag{3.1}$$

Suppose initially the frequencies of genotypes A, B, C are in the ratio p^2, $2pq$, q^2, respectively.

The following table shows the adult frequencies across the same generation after the survival rate of each genotype has been factored in.

Table 2

	$\underline{A}$	$\underline{B}$	$\underline{C}$
Juvenile Frequencies	p^2	2pq	q^2
Relative survival rates	w_A	w_B	w_C
Relative adult frequency	$\mathrm{p}_t^2 w_A$	$2\mathrm{p}_t q_t w_B$	$\mathrm{q}_t^2 w_C$
Adult frequencies	$\frac{p_t^2 w_A}{w_t}$	$\frac{2p_t q_t w_B}{w_t}$	$\frac{q_t^2 w_C}{w_t}$

Then, using Table 2, the adult frequency of p_{t+1} is

$$\begin{aligned} p_{t+1} &= p_A + \frac{1}{2}p_B \\ &= \frac{p_t^2 w_A}{w_t} + \frac{\frac{1}{2}2p_t q_t w_B}{w_t} \\ &= \frac{p_t(p_t w_A + q_t w_B)}{w_t} \\ &= \frac{p_t\left[p_t w_A + (1-p_t)w_B\right]}{w_t} \end{aligned}$$

The resulting difference equation models the change in the frequency of allele A from generation t to generation $t+1$.

$$p_{t+1} = \frac{p_t^2 w_A + p_t(1-p_t)w_B}{w_t} \tag{3.2}$$

If we assume the relative survival rates of each genotype to be $w_A = 1 - s$, $w_B = 1$, and $w_C = 1 - r$ then, the values of s and r can be positive or negative. However, $w_A, w_C \geq 0$. This implies that $r, s > 1$ (but both not zero).

Under these assumptions, we substitute equation 3.2 in 3.1and we have:

$$w_t = p_t^2 w_A + 2p_t q_t w_B + q_t^2 w_C$$

$$= p_t^2(1-s) + 2p_t q_t(1) + q_t^2(1-r)$$

$$= p_t^2 - p_t^2 s + 2p_t q_t + q_t^2 - q_t^2 r$$

$$= (p_t^2 + 2p_t q_t + q_t^2) - p_t^2 s - (1-p_t)^2 r$$

$$w_t = 1 - p_t^2 s - (1-p_t)^2 \tag{3.3}$$

Using substitution into equation 3.2 of equation 3.3 we get:

$$\begin{aligned} p_{t+1} &= \frac{p_t^2 w_A + p_t(1-p_t)w_B}{w_t} \\ &= \frac{p_t^2(1-s) + p_t(1-p_t)}{1-p_t^2 s-(1-p_t)^2 r} \\ &= \frac{p_t^2 - p_t^2 s + p_t - p_t^2}{1-p_t^2 s-(1-p_t)^2 r} \\ &= \frac{p_t(1-p_t s)}{1-p_t^2 s-(1-p_t)^2 r} \end{aligned}$$

Let

$$f(p_t) = \frac{p_t(1-p_t s)}{1-p_t^2 s-(1-p_t)^2 r}$$

Thus, for the first-order difference equation

$$p_{t+1} = f(p_t) \tag{3.4}$$

we find the equilibrium solutions by solving the following equation:

$$f(\overline{p}) = \overline{p}$$

.

$$\overline{p} = \frac{\overline{p}(1-\overline{p}s)}{1-\overline{p}^2s-(1-\overline{p})^2r}$$

$$\overline{p}[1-\overline{p}^2s-(1-\overline{p})^2r] = \overline{p}(1-\overline{p}s)$$

$$\overline{p}[1-\overline{p}^2s-(1-\overline{p})^2r-(1-\overline{p}s)] = 0$$

$\overline{p} = 0$ and $(s+r)\overline{p}^2 - \overline{p}(2r+s) + r = 0$

By applying the quadratic formula, we have:

$$\overline{p} = \frac{2r+s \pm \sqrt{4r^2+4rs+s^2-4rs-4r^2}}{2s+2r}$$

$$= \frac{2r+s \pm \sqrt{s^2}}{2s+2r}$$

$$= \frac{2r+s \pm s}{2r+2s}$$

Therefore, the three equilibrium solutions of the difference equation 3.4 are:

$\overline{p} = 0$, $\overline{p} = 1$, and $\overline{p} = \frac{r}{r+s}$.

When $\overline{p} = 0$, only the recessive allele, a, is present in the population, when $\overline{p} = 1$ only the dominant allele, A, appears in the population, and when $\overline{p} = \frac{r}{r+s}$ both dominant and recessive alleles, A and a, exist in the population.

Theorem 3.2. *Assume f' is continuous on an open interval I containing $\overline{x}$ and that $\overline{x}$ is a fixed point of f. Then $\overline{x}$ is a locally asymptotically stable equilibrium of the difference equation $x_{t+1} = f(x_t)$ if $|f'(\overline{x})| < 1$ and unstable if $|f'(\overline{x})^{|} > 1$.*

To determine the stability of each of these three equilibrium solutions, we must first find the derivative of $f(p) = \frac{p(1-ps)}{1-p^2s-(1-p)^2r}$ to be able to apply theorem 2.

$$f'(p) = \frac{(1-2ps)(1-p^2s-r+2pr-p^2r)-(p-p^2s)(2ps+2r-2pr)}{(1-p^2s-r+2rp-rp^2)^2}$$

After simplifying, the derivative is:

$$f'(p) = \frac{(1-s)p^2+2(1-s)(1-r)p(1-p)+(1-r)(1-p)^2)}{(1-p^2s-r+2rp-rp^2)^2}$$

(1) If $\overline{p} = 0$ then

$$|f'(\overline{p})| < 1 \Rightarrow f'(0) = \frac{1-r}{(1-r)^2} = \frac{1}{1-r} < 1 \Leftrightarrow 1-r > 1 \Leftrightarrow r < 0$$

So $\overline{p} = 0$ is locally asymptotically stable when $r < 0$. This is the case when the relative survival rate of genotype C is $w_C = 1 - r > 1$.

(2) If $\overline{p} = 1$ then

$$|f'(\overline{p})| < 1 \Rightarrow f'(1) = \frac{1-s}{(1-s)^2} = \frac{1}{1-s} < 1 \Leftrightarrow 1-s > 1 \Leftrightarrow s < 0$$

Therefore, $\overline{p} = 1$ is locally asymptotically stable if $s < 0$. This is the case when the relative survival rate of genotype A is $w_A = 1 - s > 1$.

(3) Lastly, if $\overline{p} = \frac{r}{r+s}$ then

$$|f'(\overline{p})| < 1 \Rightarrow f'(\frac{r}{r+s}) = \frac{2rs-r-s}{rs-r-s} < 1 \Leftrightarrow r+s-2rs < r+s-rs \Leftrightarrow rs > 0$$

Further, for this point to be stable we have to have $rs - r - s < 0$. Therefore, the values of r and s must both be positive.

So, $\overline{p} = \frac{r}{r+s}$ is locally asymptotically stable if $r, s \in (0,1)$. When $r, s \in (0,1)$, the heterozygous genotype has the largest survival rate. In this case, $w_B > max\,\{w_A, w_C\}$, the heterozygote genotype has an advantage in variability because both alleles are present.

So, mean fitness increases over time until equilibrium is reached at one of the three stability points.

Example 3.3. *The mean fitness of the population genetics model discuss before is* $w_t = p_t^2 w_A + 2p_t(1-p_t) + (1-p_t)^2 w_C$*, where* p_t *is the proportion of the population carrying allele A. Selection is governed by the survival rates* $w_A = 1 - s$ *and* $w_C = 1 - r$*, where* $r, s < 1$*. Let show that*

$$w_{t+1} - w_t = \frac{p_t(-1+p_t)[p_t^2(r+s) + p_t(s-3r) + 2r - 2][(p_t(s+r) - r]^2}{w_t^2}$$

and

(1) $w_{t+1} - w_t \geq 0 \ \textit{if}\ p_t \in [0,1]$

(2) $w_{t+1} - w_t = 0 \Leftrightarrow p_t = 0,\ \textit{or}\ p_t = 1,\ \textit{or}\ p_t = \frac{r}{r+s}$

This means that selection leads to an increase in the mean fitness of the population over time.

Proof. Let

$$p_{t+1} = \frac{p_t(1-p_ts)}{1-p_t^2s-(1-p_t)^2r}$$

according with equation 3.4 we have

$$p_{t+1} = \frac{p_t(1-p_ts)}{w_t} \tag{3.5}$$

Let find

$$w_{t+1}-w_t = [p_{t+1}^2(1-s)+2p_{t+1}(1-p_{t+1})+(1-p_{t+1})^2(1-r)]-[(p_t^2(1-s)+2p_t(1-p_t)+(1-p_t)^2(1-r)]$$

$$= [p_{t+1}^2(-s-r)+2rp_{t+1}+(1-r)]-[p_t^2(-s-r)+2rp_t+(1-r)]$$

$$= [p_{t+1}^2(-s-r)+2rp_{t+1}]-[p_t^2(-s-r)+2rp_t]$$

Substituting for p_{t+1} from equation 3.5, we get

$$w_{t+1}-w_t = \left[\frac{p_t^2(1-s)+p_t(1-p_t)}{w_t}\right]^2(-s-r)+2r\left[\frac{p_t^2(1-s)+p_t(1-p_t)}{w_t}\right]-\left[p_t^2(-s-r)+2rp_t\right]$$

$$= \left[\left(\frac{-p_t^2s+p_t}{w_t}\right)^2(-s-r)+2r\left(\frac{-p_t^2s+p_t}{w_t}\right]-\left[p_t^2(-s-r)+2rp_t\right]$$

By finding a common denominator we arrive at

$$= \left[\frac{(-p_t^2+p_t)^2(-s-r)}{w_t^2}\right]+2r\left[\frac{w_t(-p_t^2+p_t)}{w_t^2}\right]+\left[\frac{w_t^2[-p_t^2(-s-r)]}{w_t^2}\right]+\left[\frac{w_t^2[-p_t^2(-2rp_t]}{w_t^2}\right]$$

By substituting $w_t = p_t^2(1-s) + 2p_t(1-p_t) + (1-p_t)^2(1-r)$ into the above equation and further simplification gives:

$$w_{t+1} - w_t = \frac{p_t(-1+p_t)[p_t^2(r+s) + p_t(s-3r) + 2r - 2](p_t(s+r) - r)^2}{w_t^2} \tag{3.6}$$

which prove the first statement in the exercise.

Further, show that $w_{t+1} - w_t \geq 0$ for$p_t \in [0,1]$, and that $w_{t+1} = w_t$ only if p_t is one of the equilibrium solutions, $0, 1,\ or\ \frac{r}{r+s}$

Let start with equation 3.5.

If $p_t \leq 0 \Rightarrow w_{t+1} - w_t \geq 0$ when $0 < p_t < 1$ and $0 < r, s < 1$.

Since $(p_t(s+r) - r)^2$ and w_t^2 are always positive, we only need to analyze the first 3 terms of the equation 3.5: $p_t,\ (-1+p_t),\ and\ p_t^2(r+s) + p_t(s-3r) + 2r - 2$.

For this let $g(p_t) = p_t^2(r+s) + p_t(s-3r) + (2r-2)$ be a quadratic function in p_t

Under the following assumptions,

$p_t > 0$ and $p_t - 1 < 0$ After examining three cases: $g(0) < 0$, $g(1) < 0$, and $g(p_t) = 0$, we find that $g(p_t) \leq 0$ has no solutions for $p_t \in (0,1)$ because $g(p_t)$ is a quadratic function. The graph of the function must be a parabola opening upwards because $r + s > 0$. All three of the above conditions cannot be satisfied at the same time.

$$w_{t+1} - w_t = 0\ \ if\ and\ only\ if\ p_t\ is\ one\ of\ the\ equilibrium\ solutions\ :$$

$$p_t = 0\ , p_t = 1\ , or\ p_t = \frac{r}{r+s}$$

Therefore, natural selection causes the mean fitness of a given population to increase as time passes. □

Example 3.4. *Genetic Example*

A well-known example of an inherited autosomal recessive disorder caused by two alleles at a single locus is Sickle Cell Disease (SCD). There are two alleles for the production of hemoglobin, β^A and β^s. If two copies of the β^A allele are inherited, the person will not have SCD. However, a person will have this disorder if they inherit two copies of the β- globin S (β^s) allele, resulting in the formation of abnormal hemoglobin molecules. Hemoglobin is the molecule in red blood cells (RBCs) that is responsible for delivering oxygen to the bodies.

When two copies of the β^s allele are inherited, red blood cells sickle, becoming crescent shaped. The sickling of RBCs results in low numbers of RBCs, known as anemia, repeated infections, repeated episodes of pain, and the premature breakdown of RBCs. Lastly, in the heterozygous case in which each allele is inherited, the person has what is a carrier for SCD and has what is referred to as Sickle Cell Trait (SCT). Because these alleles interact codominantly, in the heterozygous case, both alleles will be expressed. This means that some RBCs will have a normal shape and some will sickle. In this case, the person is generally healthy. They may experience some symptoms of SCD, but they will be less severe.

Table 3

$\beta^A\beta^A$	$\beta^A\beta^s$	$\beta^s\beta^s$
No SCD	SCT	SCD
Normal RBCs	Some Normal RBCS, Some sickled RBCs	Sickled RBCs

Since there are only two alleles at a single locus that determine Sickle Cell Disease, the principles of Hardy-Weinberg can be applied. Given data, it could be proved that allelic and genotypic frequencies will not fluctuate from generation to generation. Thus, the occurrence of each of the three genotypes will remain constant in each subsequent generation.

References

[1] Allen, L. J. S. *An introduction to mathematical biology*, NJ: Pearson/Prentice Hall, 2007.

[2] Brger *The mathematical theory of selection, recombination, and mutation*, *Wiley, 2000*

[3] *Sickle Cell Disease U.S. National Library of Medicine* 2015, April 28