1-1-1986

Selection in genetics : a mathematical model

Sarah J. Blean

"Selection in Genetics: A Mathematical Model"

Sarah J. Blean

May 7, 1986

There are many areas of study in the field of genetics which continually affect our highly technological society. In particular, the idea of being able to predict the time when, given the current distribution of a population, a specific inherited trait might disappear from the population is one important question that is being raised. In order to accurately make such a prediction, a mathematical model will be built. Since there are several variables that enter into such a model, certain assumptions will initially be made. After studying the original model, a selected group of the variables and their corresponding assumptions will be changed to look at other possible cases. One of the key assumptions used in the first model is that of random mating. Random mating is a very broad term and excludes such things as mutations, inbreeding, selection based on genotype and inter-generation breeding.

For the basic model, define AA,aa,Aa to be the three possible genotypes. The first two are called homozygotes, and the last one is called a heterozygote.[1] If we say that the heterozygote will always show the same characteristic as the AA homozygote, then we call A the dominant gene, and a the recessive gene. In a situation where a constant, s, of the zygotes Aa and AA survive to the age where mating takes place, but only (1-k)s of the aa type zygote survive to a similar age for $0<k<1$, we would call k the selective disadvantage associated with aa. To make it easier to measure

the number of zygotes at any given generation, define the gene frequency of a particular type of gene to be the ratio of the number of that specific gene to the total number of genes present in the population. Since the population we are dealing with consists of only two types of genes, let p be representative of the gene frequency for the A genes, and let q stand for the gene frequency of the a gene, where $p + q = 1$.[2] We may now restate the purpose of the model in terms of the gene frequency q. Given an initial population where q_o is greater than some q, we would like to predict the time when q_n is less than that same q.

In our first basic model we would like to develop a recurrance relationship in order to study the gene frequencies at each generation. Using this relationship, we should be able to, given only the status of the initial population, find the desired q_n.

From the Hardy-Weinberg Ratio,[3] the nth generation gene distribution is in the form $p_n^2 : 2p_nq_n : q_n^2$. In addition, if we take into account the selective disadvantage, these ratios become $p_n^2 : 2p_nq_n : (1-k)q_n^2$. Taking a specific case where the genotypes AA, Aa, aa occur in the ratio P:Q:R where $P + Q + R = 1$, we can find that the gene frequency $p = P+Q/2$, and $q = R+Q/2$. Rewriting our gene distribution in the form of $hp_n^2 : 2hp_nq_n : h(1-k)q_n^2$, where the sum of these numbers is one, allows us to find formulas for p_{n+1} and q_{n+1}. P takes the form hp_n^2, Q that of $2hp_nq_n$, and R equals $h(1-k)q_n^2$, so that $p_{n+1} = hp_n^2 + hp_nq_n$, and $q_{n+1} = h(1-k)q_n^2 + hp_nq_n$. Now to get rid of the added term h, divide these

two quantities and let:

$$y_{n+1} = \frac{p_{n+1}}{q_{n+1}} = \frac{y_n(y_{n+1})}{y_n+1-k} \text{ , where } y_i = \frac{p_i}{q_i}$$

As was mentioned before, if we know q_o and k, we can say that $y_o = \frac{1-q_o}{q_o}$. Using our recurrance relation to find y_1, y_2, ... it turns out that $y_n = \frac{1-q_n}{q_n}$. Now use $q_n = \frac{1}{1+y_n}$ to compute the gene frequency of aa at each generation.

This procedure works very well, but it turns out that this is a very slow process and takes thousands of calculations, which could only be called efficient if a computer were correctly programmed to make these calculations. In order to make the model more readily usible, we will develop acomputer-free form of the model. This new model will replace the recurrance relation using the y_n's,with a function f, that is defined for all $x>0$ such that $f(x) \doteq y_n$ whenever n=x.

From the recurrance relation we have that:

$$y_{n+1} - y_n = \frac{ky_n}{y_n+1-k}$$

If the approximation of y_n is to be true, $f(n+1) - f(n) \doteq \frac{kf(n)}{f(n)+1-k}$ (*) must also be true. This is only for integer values of n, however, and the original intent was for f to hold for all values x 0. Replacing the n's in (*) with x's, we get $f(x+1) - f(x) \doteq \frac{kf(x)}{f(x)+1-k}$.

Since $[f(x+1) - f(x)] = [f(x+1) - f(x)]/[(x+1)-x]$, we can call this the slope of the line segment between two points $(x,f(x))$ and $((x+1),f(x+1))$ on the graph of f.[4]

For a small selective disadvantage k, f(n+1)-f(n) is small. This means that f is a rather flat curve and to closely approximate the slope of this curve we may use the tangent slope. Since the tangent slope of a curve is given by the derivative, when it exists, we assume f is differentiable and write $f'(x) \doteq \frac{kf(x)}{f(x)+1-k}$. We would like to solve this equation for f:

$$f'(x)\ \frac{f(x)+1-k}{f(x)} = k$$

$$f'(x)+(1-k)\ \frac{f'(x)}{f(x)} = k$$

$$\int f'(x)dx+(1-k)\int \frac{f'(x)}{f(x)}dx = \int k dx$$

$$f(x)+(1-k)\ln f(x) = kx+C$$

Putting x = 0 and solving for C we obtain:

$$C = f(0)+(1-k)\ln f(0)$$

$$f(x)-f(0)+(1-k)\ln\frac{f(x)}{f(0)} = kx$$

Remembering the initial substitution $f(x)=y_n$, we can say that

$$y_n-y_0+(1-k)\ln\frac{y_n}{y_0} \doteq kn \quad (**)$$

Going back to our original intent of finding out at which generation we are satisfied with a small enough q_n, we solve (**) for n:

$$n \doteq {}^1/_k[y_n-y_0+(1-k)\ln\frac{y_n}{y_0}]$$

Using both the recurrant relation, and the derived computer-free processes, computations were done using an initial q_0 and a given fixed selective disadvantage, k. The results using the approximation for n were found to be well within one-half of one percent of the true n found by using the first method.[5]

Now that we have structured a basic model for the problem,

let us look at some of the assumptions of the model that we might wish to vary. First of all, let us examine the selective disadvantage which was attatched to the recessive genotype, aa. This selective disadvantage is one of the most important parts of the model. Now we will examine it more carefully. It is most often expressed in the form of a percentage, i.e. k = 0.1 means that aa has a selective disadvantage of 10%. We would have had no reason to set up this model if the recessive genotype aa did not have a decreased chance of surviving to adulthood, in comparison with the other two genotypes. In our basic model, k was taken to be a fixed small positive number, Perhaps it might make more sense if k was not a constant, but rather if it changed over a period of time of generations, becoming worse as time goes on. So now we need to find a function, f, that could be used to replace the k in the earlier computations. The values should be decreasing over time, and in order to make sense for the model, the values should lie in the interval zero to one. Keeping the same initial q_0 that was used to test the first model, I started to try different possibilities for the desired function. Since we are dealing with a very slow process over a long period of time, f(n) might be in the form of an exponential. The exponential function has many interesting and unique properties. Letting $f(n) = e^{-\frac{n}{10000}}$, and replacing k with f(n) in the equation in which we solved for n, results in:

$$n \doteq \frac{1}{f(n)}[y_n - y_0 + (1-f(n))\ln y_n/y_0] \quad (++)$$

In terms of $f(n) = e^{-\frac{n}{10000}}$, (++) becomes:

$$n \doteq \frac{1}{e^{-\frac{n}{10000}}} [y_n - y_0 + (1 - e^{-\frac{n}{10000}}) \ln y_n/y_0]$$

Now we have a more complicated equation in which n is involved on both sides of the equation. Again we would like to solve for n, and from our new equation for n we get:

$$ne^{-\frac{n}{10000}} \doteq (y_n - y_0) + (1 - e^{-\frac{n}{10000}}) \ln y_n/y_0$$

$$n \doteq e^{-\frac{n}{10000}}(y_n - y_0) + e^{-\frac{n}{10000}} \ln y_n/y_0 - \ln y_n/y_0$$

$$\frac{n}{10000} e^{-\frac{n}{10000}} \doteq (y_n - y_0) + \ln y_n/y_0$$

This equation is in the form of $xe^{-x} = c_0$, and can be solved numberically. The results are shown below:

With $q_0 = .99$ and therefor $y_0 = .01$, and the desired $q_n = .001$ with $y_n = 999$, we can solve the equation and find that x=3.5626. But $x = \frac{n}{10000}$, so the solution for n is 35,626 generations.

Next we examine the assumption of random mating. Taking the human population as an example, derivations from random matin are quite common.[6] Marriages between close relatives seem to take place more often than could be termed as random, citing the European royalty as one well-known case of this. Another case related to non-random mating deals with the tendency humans have to pick their mates according to physical traits: Tall people are more likely to marry tall people, and redheads are <u>less</u> likely to marry redheads than could be considered random.[7] The special case which will be looked into further shows that selection based on genotype is a strong case against random mating. Specifically, the

type of selection we are interested in involves the different survival rates of individuals to the age where mating takes place. If we let this type of selection be based on genotype, it would seem reasonable that given an initial population: $p^2AA+2pqAa+q^2aa$, the survival rates might be of the form $1-S:1:1-s$, ($0 \leq S, s \leq 1$). We are then interested in the distribution of the population which reached the age of mating:

$$\frac{(1-S)p^2AA + 2pqAa + (1-s)q^2aa}{1 - Sp^2 - sq^2} \quad 8$$

The denominator of this equation is the total population, which is derived in the following manner:

$$\begin{aligned} & (1-S)p^2 + 2pq + (1-s)q^2 \\ &= p^2 - Sp^2 + 2pq + q^2 - sq^2 \\ &= p^2 + 2pq + q^2 - Sp^2 - sq^2 \\ &= (p+q) - Sp^2 - sq^2 \\ &= 1 - Sp^2 - sq^2 \end{aligned}$$

In terms of each of the gamates, A and a, then each partner contributes one gene from the array:

$$\begin{aligned} & \frac{((1-S)p^2+pq)A + ((1-s)q^2+pq)a}{1 - Sp^2 - sq^2} \\ = & \frac{(p^2-p^2S+p(1-p))A + (q^2-sq^2+(1-q)q)a}{1 - Sp^2 - sq^2} \\ = & \frac{(p-p^2S)A + (q-q^2s)a}{1 - Sp^2 - sq^2} \end{aligned}$$

Then the gene frequency, p_1, for the A gene at the birth of the next generation is:

$$\frac{p - Sp^2}{1 - Sp^2 - sq^2}$$

To determine when such a system would become stabilized, take $p_1 - p$: $= \frac{-pq(Sp-sq)}{1-Sp^2-sq^2} = -pq[(S-(S+s)q)/(1-Sp^2-q^2)]$

This is in equilibrium when : $q_0 = S/(S+s)$.

A special case of this survival rate example is when $S = 0$ and $s = 1$;[9] this means that the recessive genotype is eliminated from the population. Now we have the gene frequency at the birth of the next generation, $p_1 = \frac{p}{1-q^2}$ which equals $\frac{1}{1-q_0}$, and $q_1 = 1-p_1 = \frac{q_0}{1+q_0}$. Again wishing to find a recurrance relation, we can see that $q_n = \frac{q_0}{1+nq_0}$. Now we would like to know when n is large enough so that $q_n <$ a fixed q. This will occur when $\frac{1}{1+[1/q_0-1]2^n} < q$ (***)

Solving for n we can see that (***) becomes:

$$1 < q + q[1/q_0-1]2^n$$

$$= \frac{1-q}{q} < [1/q_0-1]2^n$$

$$= \ln(1-q)/q < 2^n \ln(1/q_0-1)$$

$$= \frac{\ln(1-q)/q}{\ln(1/q_0-1)} < 2^n, \quad \text{for } q_0 \quad 1/2$$

$$= \frac{\ln[\ln(1-q)/q/(\ln(1/q_0-1)]}{\ln 2} < n$$

Another case which is interesting is that when the heterozygote does not have the capability to reproduce.[10] Now we have the gamatic array coming only from the two homozygotes. The mating population consists strictly of

$$\frac{P^2AA + R^2aa}{P^2+R^2}, \quad \text{for } P+R = 1$$

This gives the gene frequency, q_1, for the a gene at the time of birth of the next generation $= \frac{R^2}{P^2+R^2}$.

Again, a recurrence relationship can be established such that

$$q_{n+1} = \frac{q_n^2}{1-2q_n+2q_n^2}$$

In general:

$$q_n = \frac{q_n^2}{p_0 2^n + q_0 2^n}$$

To see what will happen to q in the long run look at the relation:

$$\lim \ln((1-q_n)/q_n)^* = 2^n \ln((1-q_o)/q_o)$$

for:

q_o 1/2, * , q_n 0

q_o 1/2, * , q_n unity

q_o = 1/2, * *, q's do not change

Therefor, $q_{o=}$ 1/2 is an unstable equilibrium.

Returning to the case of S=0 and s=1, where q_o = S/S+s, we would like to study the type of equilibrium that this represents. Given a q_o S/S+s, we know that q_1 q_o, and also, given q_o S/S+s we can say that q_1 q_o. The interesting question is to ask if these sequences, $q_o, q_1 \ldots$ converge, and if so, do they converge to the equilibrium, S/S+s.

Our recurrance relation is $q_{n+1} = \frac{q_n - sq_n^2}{1-Sp_n^2-sq_n^2}$. If this relation has a limit, call it x, then:

$$x = \frac{x-x^2 s}{1-S(1-x)^2-sx^2} \quad ++$$

Letting $f(x) = \frac{x-x^2s}{1-S(1-x)^2-sx^2}$, and looking at $f'(x)$ at the point $x = 0$; we can see that because this equation has no real solutions, $f(x)$ is a function of x that is always increasing or decreasing. From (++), if we solve for x, we have that x can only be equal to 1 or to $\frac{S}{S+s}$. Now we can say that $\{q_n\}$ does have a limit, it is a strictly decreasing sequence, it is bounded on $[0,1]$, so therfor, $\{q_n\}$ does converge. But what does it converge to?

For q_0 1, and q_0 $\frac{S}{S+s}$, we do have that $q_0, q_1, \ldots$ is converging to $\frac{S}{S+s}$, so yes this is a stable equilibrium.

To determine a ratio of the given frequencies of the genes, $y_n = p_n/q_n$, we have that:

$$\frac{p_{n+1}}{q_{n+1}} = \frac{p_n - Sp_n^2}{q_n - q_n^2 s} = \frac{p_n/q_n - S p_n/q_n * p_n}{1-q_n s}$$

And therefor:

$$y_{n+1} = \frac{y_n - S y_n(1-q_n)}{1-q_n s}$$

This is not as easy to work with as the ratio y_n was in the basic model. The calculations are much more complex and a direct result of the modifications from the idea of random mating.

It was mentioned early on that random mating was a major assumption of the model, and it turns out that random mating can greatly simplify the problems of looking at long term relations in genetics. There are many other possibilities to look at in terms of the basic model and all of the ways to vary it. We have touched upon only a very few of the interesting questions dealing with"Selection in Genetics".

NOTES

1. "Selection in Genetics", UMAP Module 70, Horelick, Brindell, and Koont, Sinan. Birkhauser Boston Inc. Cambridge, 1980, p 285.

2. Horelick and Koont, p. 285

3. Horelick and Koont, p. 286

4. Horelick and Koont, p, 290

5. Horelick and Koont, p. 292

6. Principles of Human Genetics, Stern, Curt. W. H. Freeman and Company, San Francisco, 1973, p. 236

7. Stern pp. 222-223

8. An Introduction to Genetic Statistics, Kempthorne, Oscar. Iowa State University Press, Ames, Iowa, 1969, p.48.

9. Kempthorne, p. 48

10. Kempthorne, p. 49