Stochastic Processes

Jacob Denson

November 1, 2023

Table Of Contents

1 History of Population Dynamics

1.1 1202: Fibonacci's Sequence . . . .

1.2 1693: Halley . . . .

1.3 1748-1761: Leonard Euler . . . .

2 Population Genetics

2.1 The Hardy-Weinberg Law . . . .

2.2 Fisher's Work . . . .

Chapter 1

History of Population Dynamics

1.1 1202: Fibonacci's Sequence

Fibonacci's sequence emerged from a problem in population dynamics, posed in his 1202 book *Liber abaci*. The problem is to determine the number of pairs of rabbits produced from a single pair of rabbits, if that pair of rabbits produces another pair of offspring each month, and new pairs of offspring being producing pairs of rabbits two months after being born. If S_n denotes the number of pairs of rabbits at the beginning of the nth month, then $S_1 = 1$, $S_2 = 2$, and more generally, the sequence satisfies the reccurence $S_{n+1} = S_n + S_{n-1}$. We thus see the emergence of the Fibonacci sequence. In 1602, Kepler independently considered the sequence, and noticed that

$$\lim_{n\to\infty} S_{n+1}/S_n = \phi = \frac{1+\sqrt{5}}{2}.$$

Thus the Fibonacci sequence grows, roughly, at a geometric rate, a trait common to population models. In 1728 Daniel Bernoulli found the exact formula for the Fibonacci sequence

$$S_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right).$$

The complete works of Fibonacci were published in the 19th century, which resulted in the sequence being named after him.

1.2 1693: Halley

In order to more effectively price annuities, Halley came up with a more sophisticated model for the population of Breslau accounting for both deaths and births. He assumed that the population of Breslau was steady, i.e. the number of population of a given age (and thus the annual number of deaths) were all fixed and independent of time. Let's let $\{P_n : n \geqslant 0\}$ denote the population of each age (so that P_0 is the number of births each year). If $\{D_n : n \geqslant 0\}$ denotes the number of people who died at age n, then $P_{n+1} = P_n - D_n$. Using demographic data, Halley was able to compute values for P_0 and the sequence $\{D_n\}$, and thus reconstruct $\{P_n\}$. In particular, he could then use this to construct the total population $\sum_n P_n$ of Breslau. The probability of an individual surviving the nth year of their life in Breslau, was then $1 - D_k/P_k = P_{k+1}/P_k$, and this could be used reasonably for the population of other similar European cities.

Before Halley, annuities were prices at a fixed rate independent of age. But Halley realized prices could be priced much more efficiently if the price varied depending on the age of a buyer. Recall that one purchases an annuity at a fixed price, then recieves a certain percentage of this annuity each year until they die. Suppose an individual of age k bought an annuity at a price P which gives one unit of currency each year, and let i be the interest rate of money, which we can assume is fixed. The expected interest normalized value of money that must be paid back R_k would then be

$$\frac{P_{k+1}}{P_k}\frac{1}{1+i} + \frac{P_{k+2}}{P_k}\frac{1}{(1+i)^2} + \cdots = \frac{1}{P_k}\sum_{j=1}^{\infty}\frac{P_{k+j}}{(1+i)^j}.$$

If the government wants this deal to be fair, the price should be equal to this sum. In the case of Breslau, Halley calculated a 20 year old would get 7.8% of their original purchase amount each year, a 50 year old should get 10.9% each year, and so on. Using the recent invention of logarithms, Halley was able to perform this tedious calculation.

Life tables also interested Huygens, Leibniz, and Bernoulli. In 1725, Abraham de Moivre simplified the calculations considerably, by noticing that the sequence $\{R_k\}$ satisfied the backwards recurrence formula

$$R_k = \frac{P_{k+1}}{P_k}\frac{1 + R_{k+1}}{1+i}.$$

For large k the values $\{R_k\}$ contain fewer terms, and are thus more easily calculatable, and then the formula above allows us to work backward.

1.3 1748-1761: Leonard Euler

In 1748, in a treatise entitled *Introduction to Analysis of the Infinite*, Euler considered various examples of exponentials and logarithms, including questions of population dynamics. In these questions, Euler assumed that the population $\{S_n\}$ at a given year satisfied the equation

$$S_{n+1} = (1 + x)S_n,$$

for some growth rate x. One then sees that

$$S_n = (1 + x)^n S_n.$$

We thus obtain geometric growth. Using this example, and assuming a growth rate of $1/30$ people per year, Euler showed that a city could expand by at least 100,000 people in 100 years, a situation which had been observed in London at the time. A devout Protestant, Euler also defended the claim that humans could descend from a population of 6. Assuming a growth rate of $1/16$ people per year, 1,000,000 people could descend from a population of 6 in 200 years. Euler states it's therefore "ridiculous to object that in such a short space of time the whole earth could not be populated beginning with a single man".

In 1760, Euler returned to the question, discussing the problem of life tables. Assuming knowledge of the number of births $\{B_n\}$ each year, and the number of deaths $\{D_n\}$ each year, and the proportion of individuals p_k that are still alive at the beginning of the kth year of their life, Euler considered the problem of finding the population of humans over time. Euler made the following additional assumptions:

- The population grows geometrically, $S_n = r^n S_0$ for some quantity $r \geqslant 1$.
- The number of births is proportional to the population, i.e. $B_n = mS_n$.
- No one lives for more than 100 years.

Let $P_{k,n}$ be the number of people of age k on year n. Then

$$P_{k,n} = q_k B_{n-k}.$$

We also have

$$B_n = mS_n = r^n mS_0 = r^n B_0,$$

so that the sequence $\{B_n\}$ grows geometrically, at the same rate as the population. Now

$$\begin{aligned} S_{100} &= \sum_{k=0}^{100} P_{k,100} \\ &= \sum_{k=0}^{100} q_k B_{100-k} \\ &= \sum_{k=0}^{100} q_k r^{100-k} B_0. \end{aligned}$$

Noticing that $S_{100} = r^{100} S_0 = r^{100} B_0/m$, substituting this into the equation above, and dividing by $r^{100} B_0/m$ on each side, we conclude that

$$1 = m \sum_{k=0}^{100} q_k r^{-k}.$$

This equation is often called *Euler's equation of demographics*.

Our next goal is to write m as a function of r, and the known quantities, which when substituted into Euler's equation, will allow us to uniquely determine r from the known quantities. We can obtain this from the equation

$$S_{n+1} - S_n = B_n - D_n,$$

using the fact that $S_{n+1} = rB_n/m$, and $S_n = B_n/m$, we conclude that

$$\frac{r-1}{m} B_n = B_n - D_n,$$

which can be rearranged to give

$$m = \frac{r-1}{1 - D_n/B_n}.$$

Thus

$$1 = \frac{r-1}{1 - D_n/B_n} \sum_{k=0}^{100} q_k r^{-k}.$$

Thus we can determine r knowing D_n and B_n for any n, as well as the survival proportions $\{q_k\}$. Once r is determined from this equation, we can determine m

via the equation above, and thus determine

$$S_n = B_n/m = B_n \frac{1 - D_n/B_n}{r - 1} = B_n \sum_{k=0}^{100} q_k r^{-k}.$$

Note that for $r = 1$, this analysis exactly recovers the result of Halley.

Euler also considered the problem of determining the distribution of ages in a population, if the total population S_n is known, and $\{q_k\}$ and m are known. Euler's equation for demographics immediately implies that we can compute r. On a year n, we see

$$P_{k,n} = q_k B_{n-k} = q_k B_n/r^k.$$

Thus the proportion of total population that has age k is

$$\frac{q_k B_n/r^k}{B_n + q_1 B_n/r + \cdots + q_{100} B_n/r^{100}} = \frac{q_k/r^k}{1 + q_1/r + \cdots + q_{100}/r^{100}}.$$

Notice this constant is *independent of n*. Thus the 'age pyramid' keeps the same shape over time, though the number of individuals of each increases geometrically at the same rate r.

Euler also reexamined the problem of computing life tables when the population increases geometrically, i.e. to calculate the quantities $\{q_k\}$ from knowledge of census data, i.e. which allows us to compute $\{S_n\}$ (data Halley wasn't able to use). The growth rate is

$$r = \frac{S_n - D_n}{S_n - B_n}.$$

If we let $D_{k,n}$ be the number of people who died at age k during year n, then these people were born in the year $n - k$, so

$$D_{k,n} = (q_k - q_{k+1}) B_{n-k} = (q_k - q_{k+1}) B_n/r^k.$$

Thus

$$q_{k+1} = q_k - \frac{r^k D_{k,n}}{B_n}.$$

For $r = 1$, we again recover Halley's formula. Using these calculations, Euler also performed calculations for the price of annuities.

In 1761, Euler's colleague Johann Peter Süssmilch published some calculations of Euler, which gave a more detailed version of Fibonacci's model. Euler assumed a couple is initially 20 years old at year zero. Euler assumes people die at the age of 40, marry at 20, and have 6 children, 2 (a boy and a girl) at 22, another two at 24, and the last at 26. If B_i is the number of births at the year $2i$, then Euler conclude that a recurrence formula exists of the form

$$B_i = B_{i-11} + B_{i-12} + B_{i-13}.$$

Consider the initial conditions $B_{-12} = 0$, $B_{-11} = 0$, $B_{-10} = 2$, and $B_i = 0$ for $-9 \leqslant i \leqslant 0$, Euler could compute the number of births over time. In a further manuscript, Euler determined that a solution of this equation of the form cr^i satisfies

$$r^{13} = r^2 + r + 1.$$

Using a table of logarithms for the computation of r^{13}, Euler noticed that $r^{13} - r^2 - r - 1 \approx 0.212$ if $r = 1.09$, and $r^{13} - r^2 - r - 1 \approx -0.142$ if $r = 1.10$. Thus there exists a solution between 1.09 and 1.10. Approximating the function by a line between these points, Euler obtain an approximate solution $r \approx 1.0960$. Using logarithms, we can use this to determine that the population doubles approximately every 15 years. Since

$$D_i = B_{i-20} \sim cr^{i-20} = B_i/r^{20} \approx B_i/6.25,$$

the number of births is thus approximately 6 times the number of deaths. We can also conclude that

$$S_i = B_i(1 + 1/r + \cdots + 1/r^{19}) = B_i \frac{1 - r^{20}}{r^{19} - r^{20}} \approx 9.59 B_i,$$

so the total population is about ten times the number of births each year.

Chapter 2

Population Genetics

The theory of population genetics was initially developed in the 1920s and 30s by Fisher, Haldane and Wright, who, after the rediscovery of Mendelian genetics, tried to formulate an evolutionary paradigm that would follow from the theory.

When Darwin published his *origin of species*, that evolution occured was not scientifically controversial. What was more controversial was *what brought about that change*. Darwin was a *gradualist*, arguing that changes in the nature of organisms were gradual and incremental. Others, like Huxley and Galton, were *saltationists*, believing changes happened in ‘jumps’. Two schools have developed from these theories.

Before the discovery of Mendelian genetics, the standard belief was in a *blending hypothesis*; the traits of descendents are obtained by blending the traits of a descendent’s parents. Under this blending theory, the variation in a populations traits would naturally tend to zero over time, leading to a homogeneous population. Since this is not observed, there must be other effects causing a deviation of the traits of descendents. But these effects would cause problems for Darwin’s theory of evolution, which requires selectively favoured parents to produce offspring that resemble them. Darwin recognised this as a major problem to his theory. Despite being a saltationist, Galton had a close relationship with Darwin, and tried to quantify the theory of gradualism, which lead him to introduce the statistical concepts of correlation and regression. A group of scientists, which we now call the biometricians, developed to understand the theory, involving scientists such as Weldon and Pearson.

In 1900, Mendelian genetics was rediscovered. Rather than a blending theory described above, Mendel proposed that the factors that control traits are controlled by certain particles (which, perhaps apocryphally, we can call a *quantization* of the blending theory), which he called *genes*, which are passed from generation to generation. In the study of pea plants with purple and white flower, Mendel postulated that each non-reproductive *somatic cell* of a pea plant possesses two genes controlling plant color (together forming the *genotype* of the cell). However, when the plant produces reproductive cells, or *gametes*, only one copy of each gene enters. When a male and a female gamete unite, each contributes a gene to the cells of a new individual. Mendel proposed that there were two variants (*alleles*) determining the color of the flowers of a pea plant, and had the further insight that one was *dominant* over the other. A pea plant would only produce white flowers if it possessed only white flower alleles of the gene; given one or two purple flower alleles, the plant would produce purple flowers. This explained why two purple flowered plants could produce a descendent with white flowers (the pair of genes in the two purple flowered plants are white and purple), but two white flowered plants could not produce a purple flowered plant.

Mendel's theory prove appealing to the Saltationist theory. But the biometricians were disinclined to believe in the Mendelian mechanism, or at least, to believe it was not important to the theory of selection and evolution. A bitter fight followed. Sadly, arguments such as Yule (1902), which showed a mathematical analysis of Mendelian mechanisms could reconcile the two theories, were ignored. We will see, paradoxically, that Darwinism's gradualist theory *relies on* Mendelism. It would be difficult to think of a process other than the Mendelian theory in which fitness differentials between genotypes lead to changes in gene frequencies and thus ultimately to evolution.

2.1 The Hardy-Weinberg Law

Let us consider a mathematical model; we have a monoecious population (any individual can act as a male or a female), and that they mate randomly, so that each member of the next generation of the population has father and mother chosen uniformly at random from the last generation. Consider a gene with two alleles, which we label A and B, and suppose each individual possesses two copies of this gene. Let us assume the proportions of AA, AB, and BB are X, $2Y$, and Z respectively. Let us consider a table

Parent Genotypes	Probability	OffSpring Genotype
AA and AA	X^2	AA
AA and AB	$4XY$	50% AA, 50% AB
AA and BB	$2XZ$	AB
AB and AB	$4Y^2$	25% AA, 50% AB, 25% BB
AB and BB	$4YZ$	50% AB, 50% BB
BB and BB	Z^2	BB

From this table, we can calculate the proportions X', $2Y'$, and Z' of an offspring having genotype AA, AB, and BB respectively. However, it will be more efficient to calculate by using symmetry. Let $a = X+Y$ denote the proportion of A alleles in the population, and $b = Y+Z$ the proportion of B alleles. Then a is the probability that the father, or the mother, gives an A allele, and b is the probability of giving a B allele. Thus the proportion X' of offspring with genotype AA is precisely the chance both father and mother give an a allele, i.e.

$$X' = a^2 = (X+Y)^2.$$

The proportion $2Y'$ of offspring with genotype AB is the chance a father gives an a allele, and the mother gives a b allele, or vice versa, i.e.

$$2Y' = 2ab = 2(X+Y)(Y+Z).$$

The proportion Z' of offspring with genotype BB is the chance both parents give b alleles, i.e.

$$Z' = b^2 = (Y+Z)^2.$$

We thus see from this that the proportions of alleles off the offspring is exactly the same as the proportions of the allelles in the original population. We can continue this process to a third generator. Let us write X'', $2Y''$, and Z'' for the proportion of third generation phenotypes. Then

$$\begin{aligned} X'' &= (X'+Y')^2 \\ &= ((X+Y)^2 + (X+Y)(Y+Z))^2 \\ &= (X+Y)^2(X+2Y+Z) \\ &= (X+Y)^2 = X'. \end{aligned}$$

Similarily, we conclude $Z'' = Z'$, and thus $Y'' = Y'$. Thus the proportion of genotypes remains stable after one generation of offspring. This is called the

Hardy-Weinberg theorem. The final proportion can be calculated by noting that the frequency of each alleles A and B is invariant beginning from the first generation, i.e. we have $X' + Y' = X + Y$ and $Y' + Z' = Y + Z$. If we let a and b denote the proportion of the population with the A allele, then $a + b = 1$, and the following theorem is true.

Theorem 2.1. *Under the assumptions stated, a population with allele frequencies a and b settles down after a single generation to a population with genotype frequencies a^2, $2ab$, and b^2.*

The result was first published independently by Hardy and Weinberg in 1908, though numerical examples were known to Castle (1903), Pearson (1904), and Yule (1906). The fundamental importance is that *without external forces*, then there is no tendency for the distribution of genotypes of a population to vary considerably. This shows why the Mendelian theory fixes Darwin's concerns with the blending theory. Of course, we must be careful once we introduce the property of *selection* (though we will later see a Mendelian theory with selection loses variability at a rate far less than a blending theory).

2.2 Fisher's Work

Research thus began to indicate that the Mendelian and Darwinian theory could be reconciled. In 1911, Fischer, a young mathematician at the time, stressed the importance of this reconciliation, and in 1918 achieved this reconciliation.

Consider a quantitative trait of individuals resulting at some locus with alleles A and B. Let $m : \{AA, AB, BB\} \to \mathbb{R}$ be the measurement of this trait (thus we assume no environmental effects for the trait). If the frequencies of the alleles are a and b, then the mean value of the trait is thus

$$\overline{m} = a^2 m(AA) + 2abm(AB) + b^2 m(BB).$$

and the variance is

$$\sigma^2 = a^2[m(AA) - \overline{m}]^2 + 2ab[m(AB) - \overline{m}]^2 + b^2[m(BB) - \overline{m}]^2.$$

To obtain a further analysis, Fischer considered a *regression model* for m, finding the function $\alpha : \{A, B\} \to \mathbb{R}$ such that

$$m(XY) = \overline{m} + \alpha(X) + \alpha(Y) + \varepsilon(XY),$$

where the mean of the square error $\varepsilon(XY)^2$ is minimized over the population. A simple calculation shows that this means that

$$\alpha(A) = a(m(AA) - \overline{m}) + b(m(AB) - \overline{m})$$

and

$$\alpha(B) = a(m(AB) - \overline{m}) + b(m(BB) - \overline{m}).$$

Note that then $a\alpha(A) + b\alpha(B) = 0$, i.e. the average value of the function α is zero. The minimum mean square error is then

$$a^2b^2(2m(AB) - m(AA) - m(BB))^2.$$

Statistics tells us we can write this as the sum of the variance σ^2 of m, and the square of the bias of the estimator, i.e. the average value of

$$\overline{m} + \alpha(X) + \alpha(Y) - m(XY),$$

which we can write as

$$\sigma_A^2 = 2ab(am(AA) + (b-a)m(AB) - bm(BB))^2.$$

We call this quantity the *additive genetic variance*, and σ^2, also denoted σ_D^2, is the *dominance variance*.

Some calculations then show that the correlation between a father and his son is $\sigma_A^2/2\sigma_D^2$,

Statistics tells us we can write the expected mean square error

$$a^2\varepsilon(AA) + 2ab\varepsilon(AB) + b^2\varepsilon(BB)$$

as $\text{bias}^2 + \sigma^2$, where

$$\text{bias} =$$

as well as the variance of the estimator

. This is an *additive model* approximation of the function m.

$\mathbb{E}$ the average value of ε is minimized.

How about the correlation between a father and their son? Our calculations will be simplified by performing a regression, trying to find the function $\alpha : \{A, B\} \to \mathbb{R}$ which minimizes (in a least squares sense with respect to the population distribution), the error in the equations

$$m(AA) \approx \overline{m} + 2\alpha(A) \quad m(AB) \approx \overline{m} + \alpha(A) + \alpha(B) \quad m(BB) \approx \overline{m} + 2\alpha(B).$$

In effect, $\alpha(A)$ approximates the contribution to m for an A allele, and $\alpha(B)$ approximates the contribution to m for a B allele. We thus pick

$$\alpha(A) = a(m(AA) - \overline{m}) + b(m(AB) - \overline{m})$$

and

$$\alpha(B) = a(m(AB) - \overline{m}) + b(m(BB) - \overline{m})$$

. Thus $a\alpha(A) + b\alpha(B) = 0$, and the mean square error is

$$\sigma_D^2 = a^2b^2(2m(AB) - m(AA) - m(BB))^2.$$

The difference between this and the variance σ^2 is

$$\sigma_A^2 = 2ab(am(AA) + (b - a)m(AB) - bm(BB)).$$

Thus one chooses α to minimize the quantity

$$\begin{aligned} &a^2\ (m(AA) - (\overline{m} + 2\alpha(A)))^2 \\ &\quad + 2ab\ (m(AB) - (\overline{m} + \alpha(A) + \alpha(B)))^2 \\ &\quad + b^2\ (m(BB) - (\overline{m} + 2\alpha(B)))^2. \end{aligned}$$

One can calculate that we set

$$\alpha(A) = a(m(AA) - \overline{m}) + b(m(AB) - \overline{m})$$

and

$$\alpha(B) = a(m(AB) - \overline{m}) + b(m(BB) - \overline{m}).$$

Note that one then has $a\alpha(A) + b\alpha(B) = 0$. The average effect of swapping B for A is then $\alpha(B) - \alpha(A)$.

It will be convenient to convert this into an ‘additive form’, i.e. finding c, and a function $\alpha : \{A, B\} \to \mathbb{R}$ such that

$$m(AA) = c + 2\alpha(A) \quad m(AB) = c + \alpha(A) + \alpha(B) \quad m(BB) = c + 2\alpha(B).$$

Thus the function α gives the ’drift’ from the quantity c for each allele possessed by a given individual. Simple linear algebra allows us to determine $\overline{m}$ and α uniquely from the three quantities $m(AA)$, $m(AB)$, and $m(BB)$. In a given population with allele frequency a and b, the mean value of m is then precisely

$$\overline{m} = c + 2\left(a\alpha(A) + b\alpha(B)\right).$$

The variance is

and the variance of the trait is

$$a^2[m(AA) - \overline{m}]^2 + 2ab[m(AB) - \overline{m}]^2 + b^2[m(BB) - \overline{m}]^2.$$

The covariance of the trait between a father and his child can also be calculated effectively by considering each possiblity. For instance, if the father has genotype AA, the chance of his child having genotype AA is precisely a (the chance the mother passes on the allele A to her child). The covariance is then calculated as

$$ab(am(AA) + (b - a)m(AB) - bm(BB))^2.$$

Let us write

$$\sigma_A^2 = 2ab(am(AA) + (b - a)m(AB) - b(BB))^2$$

and

$$\sigma_D^2 = a^2b^2(2m(AB) - m(AA) - m(BB))^2.$$

Then $\sigma^2 = \sigma_A^2 + \sigma_D^2$, and the correlation between father and son is precisely

$$\frac{1}{2}\frac{\sigma_A^2}{\sigma_A^2 + \sigma_D^2}.$$

The correlation coefficient by dividing the covariance by the variance is then

$$ab(am(AA) + (b - a)m(AB) - bm(BB))^2/\sigma^2$$