Simulating Population Genetics*

First case study in computing applications

today: background in simple genetics

then: Random #'s

: background in probability,
  Statistics, interpreting results

then: your solutions

Simulation is often used to develop a better understanding of complicated phenomena — like population genetics

we want to learn

the tools

the methodology

* after Durand
Basic biology

chromosome \rightarrow \text{gene} \quad \text{encodes a protein}

genes have variants, called alleles

diploid organisms have 2 copies of each chromosome, hence 2 copies of each gene.

Suppose we have 2 possible alleles at a given site, say A, a

then there are 3 possible genotypes:

AA
Aa \quad \text{(same as aA)}
aa

important questions: how do frequencies of these genotypes & of A, a vary over time in a population?
Why study this?

- Agriculture - breeding better crops & livestock
- Understanding nature of new species
- Understanding propagation of genetic diseases in populations
- Understanding history of evolution - human migration, human diversity etc.

alternatives

- Simulation
- Field experiment
- Mathematical model
Field Experiment

- Very good picture of one particular situation
- No need to abstract/approximate
- Hard work
- Difficult to measure
- Needs analysis to make predictions

Mathematical Model

- Can yield simple intuitive explanations
- Requires simplifying assumptions
- Mathematics not always tractable
- Often models aggregate or average quantities rather than snapshots

Simulation

- Can show time variation over many generations, spatial variations
- Works even when math is intractable
- Requires different or fewer simplifying assumptions
- Needs analysis to make predictions
A Very Simple Model of Population Genetics

Context: 19th century biologists (including Darwin) believed in some form of blending inheritance—offspring inherit characteristics that are an intermediate mixture of parental characteristics.

\[\implies\text{variation in population decreases requires unreasonable mutation rate to account for selection/evolution.}\]

Circa 1914: Hardy-Weinberg (Tschetverikov) Equilibrium—shows how diversity is maintained by Mendelian laws.

Sources:
- [GMSetal 93]
- [Smi 86]
- [Smi 89]
- [EK88]—more mathematical
- [EWE79]—very mathematical
- [Cro 88]—ultraselfish genes
Simplest Model

- 1 locus, 2 alleles A, a
- infinite population
- random mating
- no mutation, migration
- all equally fit

At Generation i

\[
\begin{align*}
\text{frequency of allele } A & \text{ in both sperm and eggs} = p \\
\text{frequency of allele } a & = q = 1 - p
\end{align*}
\]

random mating is equivalent to randomly matching sperm and egg ...

\[
\begin{align*}
\text{prob } \{ AA \} & = p^2 \\
\text{prob } \{ Aa \} & = 2pq \\
\text{prob } \{ aa \} & = q^2
\end{align*}
\]

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>prob</td>
<td>$p^2$</td>
<td>$2pq$</td>
<td>$q^2$</td>
</tr>
</tbody>
</table>

Hardy-Weinberg equilibrium $\Rightarrow$ at Generation i+1
Notice that if we start at generation $i$ out of equilibrium—
we reach H-W equil. in one generation.

**Example**

\[
\begin{align*}
\text{Generation } i & \quad \begin{cases} 
\text{half } AA \\
\text{half } aa \\
\text{no hetero zygotes}
\end{cases}
\end{align*}
\]

then at Generation $i$, $p = \frac{1}{2}$, $q = \frac{1}{2}$

at Generation $i+1$,

\[
\begin{align*}
\text{prob. } AA & = \frac{1}{4} \\
\text{prob. } Aa & = \frac{1}{2} \\
\text{prob. } aa & = \frac{1}{4}
\end{align*}
\]

$\Rightarrow$ then stays that way forever.

---

**FIG. 3.2.** A geometrical representation of the Hardy-Weinberg ratio.

Maynard Smith

[Smi 89]
Suppose I take $R$ measurements of something (say coin weights): $x_1, x_2, \ldots, x_R$

**mean** $= \frac{1}{R} \sum_{i=1}^{R} x_i = \bar{x}$

$V_x = \text{Variance} = \frac{1}{R} \sum_{i=1}^{R} (x_i - \bar{x})^2 \quad \rightarrow \text{mean-square deviation}$

$= \frac{1}{R} \sum_{i=1}^{R} (x_i^2 - 2x_i \bar{x} + (\bar{x})^2)$

$= \frac{1}{R} \sum_{i=1}^{R} x_i^2 - 2 \bar{x} \sum_{i=1}^{R} x_i + (\bar{x})^2 \frac{\sum_{i=1}^{R} x_i}{R}$

$= \frac{1}{R} \sum_{i=1}^{R} x_i^2 - 2 (\bar{x})^2 + (\bar{x})^2$

$V_x = \frac{1}{R} \sum_{i=1}^{R} x_i^2 - (\bar{x})^2$

These are results of experiments - should properly be called **sample mean**, **sample variance**.
Important Application: the Wahlund Effect

Suppose a population is subdivided into subpopulations

\[ \begin{array}{ccc}
  \text{freq. } A &=& \frac{1}{P_i} \\
  \text{freq. } A &=& \frac{1}{P_2} \\
  \text{freq. } A &=& \frac{1}{P_k}
\end{array} \]

There are \( p_i^2 \) homozygote AA in population 1
\( \ldots \)
\( \ldots \)
\( \ldots \)
\( \ldots \)
\( \ldots \) etc.

mean value of AA frequency in population as a whole
\[ \frac{1}{k} \sum_{i=1}^{k} p_i^2 \]

average frequency of A over whole population
\[ \bar{p} = \frac{1}{k} \sum_{i=1}^{k} p_i \]

By previous result
\[ \text{freq. of AA} = V_p + (\bar{p})^2 \quad (V_p = \frac{1}{k} \sum p_i^2 - (\bar{p})^2) \]

But Hardy-Weinberg frequency would be \((\bar{p})^2\)

\[ \Rightarrow \text{deficiency of heterozygotes in sample of segregated population} \]
Hardy-Weinberg ratio suggests models for deviation →
non-random mating? differential viability? segregated populations?

**Example [Smi89]**

Da Cunha's data on *Drosophila polymorpha*:

Abdomen color determined by single gene, two alleles E, e
EE → dark   Ee → intermediate   ee → pale

Collected 8070 flies in Brazil:

<table>
<thead>
<tr>
<th></th>
<th>EE</th>
<th>Ee</th>
<th>ee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed (O):</td>
<td>3969</td>
<td>3174</td>
<td>927</td>
</tr>
</tbody>
</table>

\[
p(E) = \frac{2 \times 3969 + 3174}{2 \times 8070} = 0.6885
\]

\[
p(e) = 1 - p(E) = 0.3115
\]

Expected from H-W ratio: \( p^2 \times 8070, 2pq \times 8070, q^2 \times 8070 \)

<table>
<thead>
<tr>
<th></th>
<th>EE</th>
<th>Ee</th>
<th>ee</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-W:</td>
<td>3825</td>
<td>3462</td>
<td>783</td>
</tr>
</tbody>
</table>

Viability? (unlikely because of lab experiments)
mating preference? (possible?)

Suggests sampling from sub-populations, each H-W
Is this deviation from theory "significant"?

This is a statistics question—we'll return to it.

Assignment: Relax two assumptions in H-W model

1) Population size finite
2) Mating not nec. random (extra?)

Hand in & report on in class:

1) Description of simulation model
   - High level
2) Results
3) Analysis
   - Qualitative
   - Quantitative
   - New hypotheses?