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

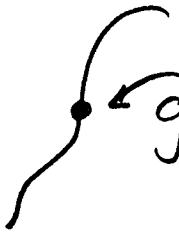
- the tools
- the methodology

---

*after Durand

Basic biology

chromosome



gene

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 (same as aA)
aa

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

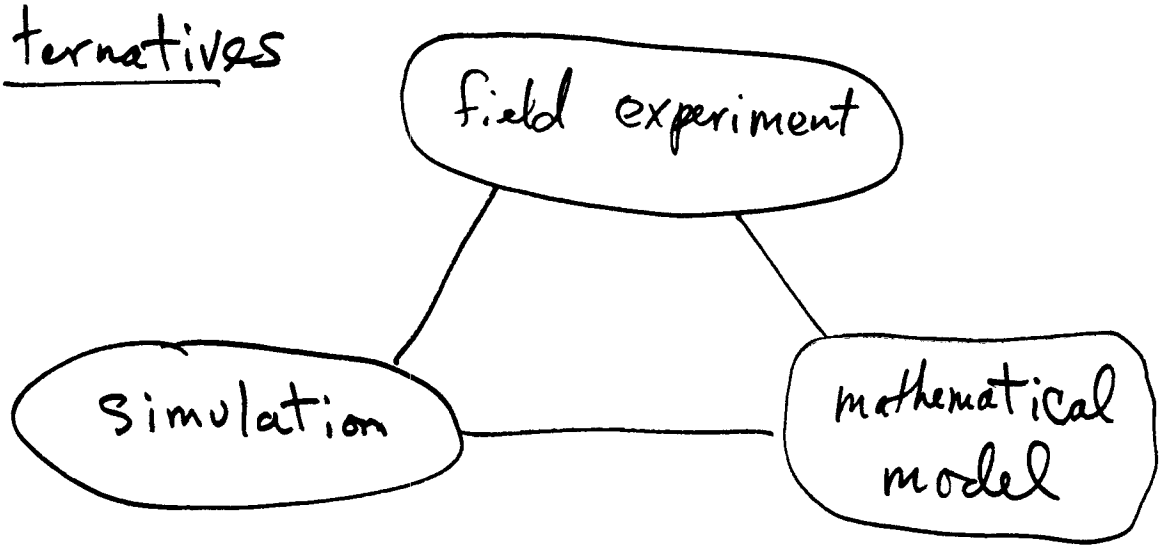
Why study this?

- 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



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 characteristic

⇒ implies that 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
- ∞ population
- random mating
- no mutation, migration
- all equally fit

At Generation i

$$\begin{cases} \text{frequency of allele A in both sperm and eggs} & = p \\ \text{frequency of allele a} & = q = 1-p \end{cases}$$

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

$$\text{prob}\{AA\} = p^2$$

$$\text{prob}\{Aa\} = pq + qp = 2pq \quad \text{(aA \& Aa are the same)}$$

$$\text{prob}\{aa\} = q^2$$

AA	Aa	aa
p^2	$2pq$	q^2

Hardy-Weinberg equilibrium → at Generation $i+1$

Notice that if we start at Generation i out of equilibrium — we reach H-W equil. in <u>one generation</u>.

<u>Example</u>

Generation i { half AA, half aa, no heterozygotes

then at Generation i, $p = 1/2$, $q = 1/2$

at Generation $i+1$, { prob.$\{AA\} = 1/4$, prob.$\{Aa\} = 1/2$, prob.$\{aa\} = 1/4$

$\Rightarrow$ then stays that way <u>forever</u>.

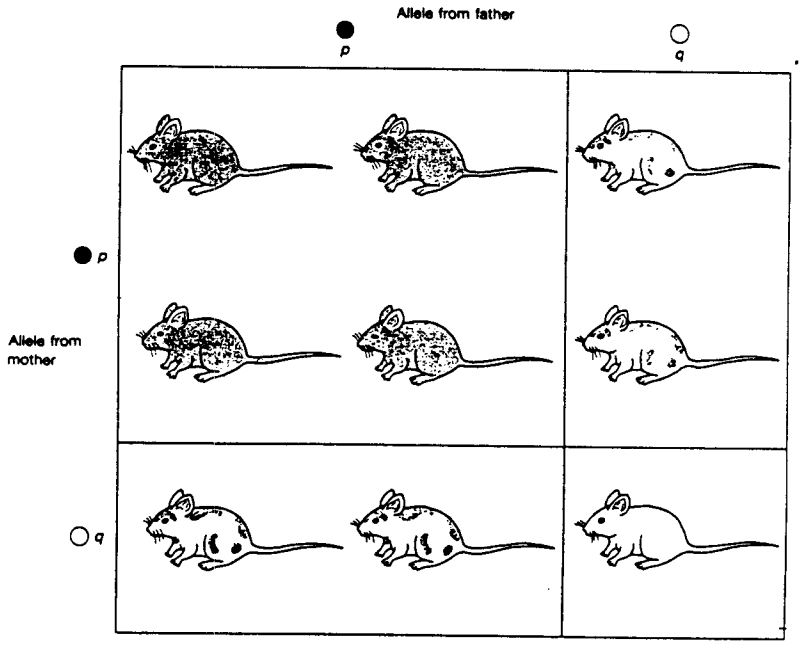


FIG. 3.2. A geometrical representation of the Hardy-Weinberg ratio.

Maynard Smith [Smi 89]

Review of Simplest Probability Theory / Statistics

Suppose I take k measurements of something (say coin weights): $x_1, x_2, \ldots, x_k$

$$\boxed{\text{mean}} = \frac{1}{k}\sum_{i=1}^{k} x_i = \bar{x}$$

$$V_x = \boxed{\text{variance}} = \frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2 \longrightarrow \text{mean-square deviation}$$

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

$$= \frac{1}{k}\sum_{i=1}^{k} x_i^2 - 2\bar{x}\,\frac{1}{k}\sum_{i=1}^{k} x_i + (\bar{x})^2 \frac{1}{k}\sum_{i=1}^{k} 1 \qquad \left(\frac{1}{k}\sum_{i=1}^{k} 1 = 1\right)$$

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

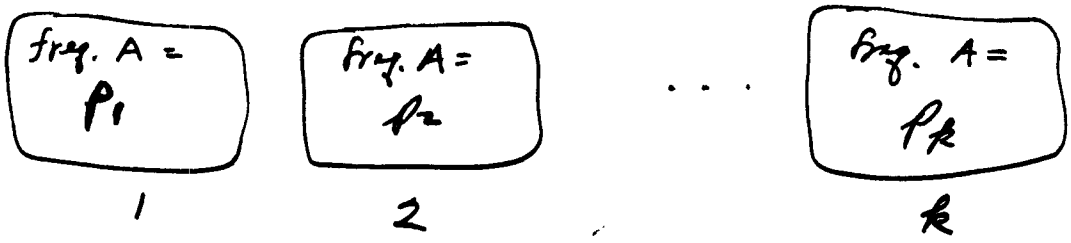
$$\boxed{V_x = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - (\bar{x})^2}$$

these are results of experiments — should properly be called sample mean, sample variance.

Important Application: the Wahlund Effect

[Smi89]

Suppose a population is subdivided into subpopulations



there are p_1^2 homozygotes AA in population 1

p_2^2 2

... 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

$$\boxed{\text{freq. of AA} = V_p + (\bar{p})^2} \qquad \left(V_p = \frac{1}{k}\{p_i^2 - (\bar{p})^2\}\right.$$

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

$\Rightarrow$ deficiency of heterozygotes in samples 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:

	EE	Ee	ee	// total
Observed (O):	3969	3174	927	8070

$$p(E) = \frac{2*3969 + 3174}{2*8070} = 0.6885$$
$$p(e) = 1 - p(E) = 0.3115$$

Expected from H-W ratio: p^2*8070, $2pq*8070$, q^2*8070

	EE	Ee	ee
H-W:	3825	3462	783

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 credit?)

Hand in & Report on in class:

1) description of simulation model
   - high level
2) results
3) analysis
   - qualitative
   - quantitative
   - new hypotheses?