Contents

Part III

Hardy Weinberg Assumption Violations

1 Nonrandom Mating

Violating Random Mating

HWE Assumption: Individuals mate at random.

Violations of this assumption can lead to HWD. We will now try to quantitate this deviation and use it to estimate quantities related to a specific type of nonrandom mating. But first, can you think of ways in which populations do not mate randomly?

- **Sexual selection.** Some genotypes may be more successful maters.
- **Disassortative mating.** Genotypes *unlike* each other have increased tendency to mate. Examples:
  - self-incompatibility in plants,
  - Roberts SC, Gosling LM, Carter V, Petrie M. (2008) MHC-correlated odour preferences in humans and the use of oral contraceptives. *Proc. Biol. Sci.*
    - single women prefer odors of MHC-similar men, women in relationships prefer odors of MHC-dissimilar men
    - significant shift to MHC-similar preference with oral contraceptive use
- **Assortative mating.** Genotypes *like* each other have increased tendency to mate.
- **Asexual reproduction.** Some fraction of the population may reproduce asexually.
- **Inbreeding.** Mating with relatives

1.1 Self Fertilization

1.1.1 Complete Selfing

Complete Selfing

Considering that all offspring of self-fertilizing homozygotes are homozygotes and 1/2 of all offspring of self-fertilizing heterozygotes are homozygotes, what do you think happens to the frequency of homozygotes in self-fertilizing populations?

$$\begin{aligned} P_{11}(t+1) &= P_{11}(t) + \frac{1}{4}P_{12}(t) \\ P_{12}(t+1) &= \frac{1}{2}P_{12}(t) \\ P_{22}(t+1) &= P_{22}(t) + \frac{1}{4}P_{12}(t) \end{aligned}$$

Focusing on allele frequencies, we have

$$\begin{aligned} p_1(t+1) &= P_{11}(t+1) + \frac{1}{2}P_{12}(t+1) \\ &= P_{11}(t) + \frac{1}{4}P_{12}(t) + \frac{1}{4}P_{12}(t) \\ &= P_{11}(t) + \frac{1}{2}P_{12}(t) = p_1(t) \end{aligned}$$

Consequences of Complete Selfing

- The proportion of heterozygotes decreases relative to HWE predictions. In fact, if there is strict self-fertilization, there will eventually be no more heterozygotes.
- The allele frequencies do not change if all other HWE assumptions are satisfied.

1.1.2 Partial Outcrossing

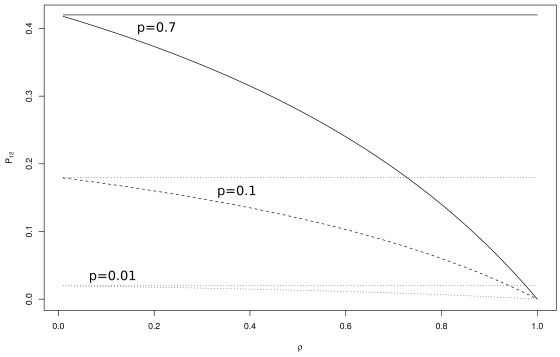
Partial Outcrossing

Many plants reproduce through a mixture of self-fertilization and cross-fertilization (aka *outcrossing*). Let ρ be the fraction of seeds fertilized by self, then

$$\begin{aligned} P_{11}(t+1) &= (1-\rho)p_1^2(t) + \rho\left[P_{11}(t) + \frac{1}{4}P_{12}(t)\right] \\ P_{12}(t+1) &= (1-\rho)2p_1(t)[1-p_1(t)] + \frac{\rho}{2}P_{12}(t) \\ P_{11}(t+1) &= (1-\rho)[1-p_1(t)]^2 + \rho\left[P_{22}(t) + \frac{1}{4}P_{12}(t)\right] \end{aligned}$$

Consider the effect on allele frequencies

$$\begin{aligned} p_1(t+1) &= (1-\rho)p_1^2(t) + \rho\left[P_{11}(t) + \frac{1}{4}P_{12}(t)\right] \\ &\quad + (1-\rho)p_1(t)[1-p_1(t)] + \frac{\rho}{4}P_{12}(t) \\ &= (1-\rho)p_1(t) + \rho P_{11}(t) + \frac{\rho}{2}P_{12}(t) \\ &= (1-\rho)p_1(t) + \rho p_1(t) \\ &= p_1(t) \end{aligned}$$



This time the proportion of heterozygotes does not decrease to 0, but it is *lower* than we would expect at HWE. If there is an equilibrium value P_{12} for which genotype frequencies no longer change, then it must satisfy the recurrence relation for heterozygotes:

$$P_{12} = (1-\rho)2p_1(1-p_1) + \frac{\rho}{2}P_{12}$$

i.e.

$$P_{12} = \frac{(1-\rho)2p_1(1-p_1)}{1-\rho/2}$$

One can also show that this equilibrium is approached, though quite slowly, as generations of partial self fertilization at constant rate ρ progress.

1.2 Inbreeding f

Inbreeding Model

The complete genotype equilibrium for the proceeding model is

$$\begin{aligned} P_{11} &= p_1^2 + \frac{\rho p_1(1-p_1)}{2-\rho} \\ P_{12} &= \frac{(1-\rho)2p_1(1-p_1)}{1-\rho/2} \\ P_{22} &= (1-p_1)^2 + \frac{\rho p_1(1-p_1)}{2-\rho} \end{aligned}$$

but these equations should look familiar if only we substituted $f = \frac{\rho}{2-\rho}$

$$\begin{aligned} P_{11} &= p_1^2 + fp_1(1-p_1) \\ P_{12} &= (1-f)2p_1(1-p_1) \\ P_{22} &= (1-p_1)^2 + fp_1(1-p_1) \end{aligned}$$

As we know, f is a measure of HWD, but now we see it as a measure of the degree of self fertilization in a population that satisfies HWE assumptions except for random mating because of partial self fertilization. In this context of non-random mating, f is known as the *equilibrium inbreeding coefficient*.

1.3 Estimating Outcrossing

Estimation of Outcrossing $1-\rho$

We now focus on methods for estimating ρ from data. An example application:

Nassar JM, Ramírez N, Lampo M, González JA, Casado R, Nava F. (2007) Reproductive biology and mating system estimates of two Andean melocacti, Melocactus schatzlii and M. andinus (Cactaceae). *Ann Bot (Lond).* **99**(1):29-38.

BACKGROUND AND AIMS: The genus Melocactus comprises 36 species of globose cacti with the most derived traits in the Cereeae tribe. It is the proper study system to examine what are the most derived reproductive strategies within that tribe. This study aims to characterize the reproductive biology and to estimate the mating system parameters of two Andean melocacti, Melocactus schatzlii and M. andinus. METHODS: The reproductive attributes of the two species were described, including floral morphology, anthesis patterns, floral rewards, floral visitors and visitation patterns. Levels of self-compatibility and autonomous self-pollination were estimated by hand-pollination experiments. Mating system estimates were obtained by conducting progeny array analyses using isozymes. KEY RESULTS: The flowers of the two species present the typical hummingbird-pollination syndrome. Despite their morphological resemblance, the two species differ in flower size, pollen and ovule production and anthesis pattern. Their main pollinator agents are hummingbirds, four species in M. schatzlii and one species in M. andinus. Both cacti are self-compatible and capable of self-pollination without the aid of pollen vectors. Population-level outcrossing rate was higher for M. schatzlii (t(m)=0.9) than for M. andinus (t(m)=0.4). At the family level, outcrossing rates for most mothers of M. schatzlii were higher (t(m)>0.8) than for M. andinus (t(m)<0.5). CONCLUSIONS: Although the two cacti are capable of selfing, M. schatzlii is a predominantly outcrossing species, while M. andinus behaves as a mixed-mating cactus. Hummingbirds are the only pollinators responsible for outcrossing and gene flow events in these species. In their absence, both melocacti set seeds by selfing. Based on its low population size, restricted distribution in Venezuela, low rates of floral visits, and high levels of inbreeding, M. andinus is considered to be an endangered species deserving further study to define its conservation status.

1.3.1 Homozygous Female Parents

Homozygous Female Parents

It is often possible to collect offspring as seeds on the mother plant so that the genotype of the mother and offspring are known. It is generally much harder to know the genotype of the pollen-contributor, i.e. the father. We will consider here estimation of ρ using data on offspring counts from mothers with known homozygous genotypes.

Consider a mother of genotype type A_uA_u. The probability that an offspring is heterozygous is the probability that an *outcross* happened and an allele not matching the mother's allele was selected.

$$P(\text{heterozygous offspring}) = \sum_{v \neq u} (1-\rho)p_v = (1-\rho)(1-p_u)$$

Since an offspring is either homozygous or not, the number of heterozygous offspring out of n_u total offspring sampled of A_uA_u mothers h_u follows a binomial distribution, with likelihood

$$L(h_u; p_u, \rho) = \binom{n_u}{h_u} \left[(1-\rho)(1-p_u)\right]^{h_u} \left[1-(1-\rho)(1-p_u)\right]^{n_u-h_u}.$$

If we observe offspring from mothers of m different homozygous genotypes, then by independence of the mothers, the total likelihood is

$$\begin{aligned} &P(h_1, \ldots, h_m; p_1, \ldots, p_m, \rho) \\ &= \textstyle\prod_{u=1}^{m} \binom{n_u}{h_u} \left[(1-\rho)(1-p_u)\right]^{h_u} \left[1-(1-\rho)(1-p_u)\right]^{n_u-h_u} \end{aligned} \tag{1}$$

This likelihood involves $m+1$ parameters (or m if m is the total number of alleles at the A locus). In general, it is not an easy likelihood to maximize. In fact, there is missing information. Can you identify it?

There are two types of homozygous offspring: (1) those resulting from selfing and (2) those resulting from fertilization by the same allele as the mother.

Aside: Complete Likelihood

If we let s_u be the number of homozygotes produced by selfing and o_u be the number of homozygotes produced by outcrossing, then the full data likelihood reduces to a simple binomial

$$\begin{aligned} P(h+o, s; p, \rho) &= \prod_{u=1}^{m} \binom{n_u}{h_u + o_u} (1-\rho)^{h_u+o_u} \rho^{s_u} \\ &= \binom{n}{h+o} (1-\rho)^{h+o} \rho^{s} \end{aligned}$$

where $h + o$ is the total number of outcrossed offspring, s is the number of selfed offspring, and $n = \sum_u n_u$ is the total number of sampled offspring. This likelihood has the trivial (and obvious) mle

$$\hat{\rho} = \frac{h+o}{n}.$$

This discussion is just meant to point out yet another situation where the EM algorithm could be helpful. For now, we will take another tact.

Homozygous Female Parents (cont.)

Suppose that the allele frequencies are known (or well-estimated, i.e. with small variance). Then the likelihood 1 involves a single parameter $\hat{\rho}$. Setting the score equal to 0 yields

$$\begin{aligned} -\sum_{u=1}^{m} \left[\frac{h_u}{1-\hat{\rho}} - \frac{(n_u - h_u)(1-p_u)}{1-(1-\hat{\rho})(1-p_u)} \right] &= 0 \\ \sum_{u=1}^{m} \frac{(n_u - h_u)(1-p_u)}{1-(1-\hat{\rho})(1-p_u)} &= \frac{h}{1-\hat{\rho}} \end{aligned}$$

The equation can not be solved explicitly for $\hat{\rho}$ unless there is only one type of homozygous mother, say AA, then

$$\hat{\rho} = 1 - \frac{h_A}{n_A(1-p_A)}.$$

General Case: Iterative Solution

In the general case, you can find a solution by repeated iteration. Plug in the current ρ_t on the right-hand side, set the left-hand side to the result and solve for the next ρ_{t+1}. Or, turn to R.

As an example, consider the following, rather obvious, data

Maternal Gene (u)	n_u	h_u	p_u
1	22	3	0.35
2	15	0	0.14
3	41	4	0.51

```
> f <- function(rho, n=n, h=h, p=p) {
   sum( (n-h)*(1-p)/(1-(1-rho)*(1-p)) ) - sum(h)/(1-rho)
}
> n <- c(22,15,41)
> h <- c(3,0,4)
> p <- c(0.35, 0.14, 0.51)
> uniroot(f=f, interval=c(0,1), n=n, h=h, p=p)
$root
```

```
[1] 0.8539492

$f.root
[1] -0.002625271

$iter
[1] 8

$estim.prec
[1] 6.103516e-05
```

Variance of $\hat{\rho}$

Of course, to make much of our estimate, we need Var($\hat{\rho}$).

- Fisher's method applies because $\hat{\rho}$ is an (implicit) function of counts (if allele frequencies are considered fixed) or counts and frequencies (if allele frequencies are estimated). However, note that if you estimate allele frequencies from the data, the counts and sample allele frequencies are correlated. (It is pretty messy.)
- Bootstrap or jackknife might just be easier. How would you resample?

  Recall that we consider each mother genotype independent. Then, within a given mother type, the heterozygotes are binomial conditional on the number of offspring n_u sampled. Thus, within each mother type, one resampled bootstrap dataset would be obtained either as

  ```
  > du <- c(rep(1,hu), rep(0,nu-hu))
  > hu.boot <- sample(du, replace=T)
  > du.boot <- length(hu.boot[hu.boot==1])
  ```

  or

  ```
  > du.boot <- rbinom(n=1, size=nu, prob=hu/nu)
  ```

1.3.2 From Hardy-Weinberg Deviations

Estimation from HWD

Of course, we could also estimate $\hat{f}$ using any of the techniques previously described and then use

$$f = \frac{\rho}{2-\rho} \implies \hat{\rho} = \frac{2\hat{f}}{1+\hat{f}}$$

As for its variance, we can use the delta method:

$$\begin{aligned} \text{Var}(\hat{\rho}) &\approx \left(\frac{\partial \rho}{\partial f}\right)^2 \text{Var}(\hat{f}) \\ &= \frac{4}{(1+f)^4}\text{Var}(\hat{f}) \end{aligned}$$

General Delta Method for Variance: For an invertible, non-zero function $T(\theta)$ that is a function of a multivariate vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, then

$$\text{Var}(T) = \sum_{i=1}^{n}\left(\frac{\partial T}{\partial \theta_i}\right)^2 \text{Var}(\theta_i) + \sum_{i=1}^{n}\sum_{j\neq i}\frac{\partial T}{\partial \theta_i}\frac{\partial T}{\partial \theta_j}\text{Cov}(\theta_i, \theta_j)$$

1.3.3 Other Estimation Related to Inbreeding/Outcrossing

Multilocus Estimates of Outcrossing

If data are available from multiple loci, then each locus gives an opportunity to identify an outcross by the presence of heterozygosity in the offspring.

Loci				
A	B	C	D	E
Maternal Genotype				
11	22	12	13	23
Offspring Genotypes				
11	22	12	13	13*
11	22	22	11	33
11	12*	12	23*	13*

Let $t = 1 - \rho$ be the probability of outcross and α be the probability of an undiscerned outcross. Then, the number of discernible outcross offspring n_o is distributed as

$$n_o \sim \text{Bin}\left(n, t(1-\alpha)\right)$$

If α is considered known, then the MLE for outcrossing proportion is

$$\hat{t} = \frac{n_o}{n(1-\alpha)}$$

Now consider α further. Suppose β_l is the probability that outcrossing cannot be detected at locus l, then

$$\alpha = \prod_l \beta_l$$

and

$$\beta_l = \sum_u P_{lu,lu} p_{lu} + \sum_u \sum_{v \neq u} P_{lu,lv}(p_{lu} + p_{lv})$$

where $P_{lu,lv}$ is the proportion of offspring sampled that have mother with genotype uv at locus l.

Estimating Number of Fathers

Suppose the mother and father genotypes are unknown and you sample offspring genotypes. Further suppose that multiple fathers could have contributed pollen to the offspring, and you want to estimate the *number* of fathers contributing paternity. Consider the following mating tables for one or two equally likely fathers.

			$P_1(O \mid M)$		
Mother	Father	Mating Prob.	AA	Aa	aa
AA	AA	p_A^4	1	0	0
	Aa	$2p_A^3 p_a$	0.5	0.5	0
	aa	$p_A^2 p_a^2$	0	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

				$P_2(O \mid M)$		
Mother	Father 1	Father 2	Mating Prob.	AA	Aa	aa
AA	AA	AA	p_A^6	1	0	0
		Aa	$2p_A^5 p_a$	0.75	0.25	0
		aa	$p_A^4 p_a^2$	0.5	0.5	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Now, suppose that the probability of one father is ϕ and two fathers is $1-\phi$. Then, the likelihood of observing the offspring genotype counts n is

$$P(n; p_A, \phi)$$

Maximize to get $\hat{\phi}$ and the estimate number of fathers is $\hat{n}_f = 2 - \phi$, rounded if you want an integer.