10 Non-random mating, Inbreeding and Population Structure.

Jewelweed, *Impatiens capensis*, is a common woodland flower in the Eastern US. You may have seen the swollen seed pods that explosively pop when you touch time, which is the source of an alternative name for this plant, "touch-me-not". Another interesting thing is that it produces two kinds of flowers.

Many of the flowers are orange and conspicuous. Those flowers have various features to attract pollinators, primarily nectar that is secreted at the base of the curved spur in the back of the flower. Bees and hummingbirds visit the flower to feed on the nectar and in the process transfer some of the pollen from one flower to another.

Most plant species produce both pollen and ovules in the same flower so there is a potential for self-fertilization to occur. That can happen either when pollen is transferred to the stigma of the same flower, or when pollen is transferred between flowers on the same plant. The showy flowers have evolved various mechanisms to limit self-fertilization within a flower. They typically produce pollen for one day, after which the anther falls off to reveal the stigma. Therefore pollen receipt (female function) is separated in time from pollen donation (male function). Within flower selfing is minimal because pollen and stigma are not functional at the same time.

In addition to the showy orange flowers, the plant produces another set of flowers that never even open. Those tiny green *cleistogamous* (meaning "closed mating") flowers look more like buds. Pollen is shed inside the unopened flower directly onto its stigma, so all of the seeds are produced by self-fertilization. Presumably the cleistogamous flowers have a reproductive assurance advantage, since they are able to produce seeds even in the absence of pollinators.

Thus *Impatiens* shows two extremes of mating: 100% self-fertilization in the tiny cleistogamous flowers, and normal crossing between plants in the showy ("*chasmogamous*") flowers.

10.1 What are the effects of nonrandom mating on allele and genotype frequencies?

The Hardy Weinberg Equilibrium was defined for the case of random mating. However it is rare that individuals in a population truly mate at random[1]. Very often populations are

[1] It is important to remember that we are talking about random mating *with respect to a particular locus*. Birds may choose mates non-randomly with respect to plumage, but since unlinked loci are inherited independently, they may very well be mating at random with respect to a certain enzyme locus.

spatially structured and individuals are more likely to mate with others that are nearby than with individuals from farther away. Often social structure will limit mating opportunities on others within the group. And some species, like jewelweed, routinely self-fertilize.

Self-fertilization: Lets start with the extreme case of self-fertilization. How do allele and genotype frequencies change after one generation of self-fertilization? As before we'll consider a single locus with two alleles, A and a. We'll use upper case letters for the genotype frequencies in that population. Let P be the frequency of AA homozygotes, H be the frequency of heterozygotes and Q be the frequency of aa homozygotes. Because those are all of the possibilities, $P+H+Q=1$.

After selfing, all of the AA homozygotes produce only AA offspring and all of the aa homozygotes produce only aa offspring. However, selfing within the heterozygotes produces all three genotypes of offspring: 1/4 will be AA, 1/2 will be Aa and 1/4 will be aa.

Genotype	Initial Frequency	Frequency after selfing
AA	P	P + 1/4 H
Aa	H	1/2 H
aa	Q	Q + 1/4 H

The result is that the frequency of the homozygous genotypes increases with each generation of selfing, and the frequency of heterozygotes decreases.

In particular, the change in heterozygosity with selfing is:

$$H' = \frac{1}{2}H$$

That shows that the frequency of heterozygotes decreases by half each generation. Eventually, after many generations of complete selfing the frequency of heterozygotes will decline to zero.

Effect of selfing on allele frequency: You can always calculate the allele frequency as $p = P + \frac{1}{2}H$. So after selfing, and substituting the new values for P and H we get

$$p' = P' + \frac{1}{2}H'$$

$$p' = P + \frac{1}{4}H + \frac{1}{2}\left(\frac{1}{2}H\right)$$

$$p' = P + \frac{1}{2}H$$

$$p' = p$$

which is simply the original allele frequency.

That is an interesting result: non-random mating, even in the most extreme form of self-fertilization, has no effect on allele frequency. Selfing causes genotype frequencies to change as the frequency of homozygotes increases and the frequency of heterozygotes decreases, but the allele frequency remains constant.

Because non-random mating only reshuffles genotype frequencies with respect to their HW expectations, we can use the deviation of genotype frequencies from their expected values as a measure of inbreeding. If H_o is the observed frequency of heterozygotes and H_e is the expected frequency under random mating, then the population inbreeding coefficient or fixation index is

$$F = 1 - \frac{H_o}{H_e}.$$

If there are no heterozygotes in the population then the fixation index is 1.0. When the frequency of heterozygotes equals the HW expectation (as with random mating) then the fixation index is 0. In cases where there is an excess of observed heterozygotes, then the fixation index can be negative.

Example: Let's assume the population starts out in HWE. That means the genotype frequencies will be p^2, 2pq and q^2. Imagine that all individuals in the population reproduce only through self-fertilization. How will the genotype frequencies change?

Genotype	Initial Frequency	Frequency after selfing
AA	p^2	$p^2 + \frac{1}{2}p(1-p)$
Aa	2p(1-p)	
aa	$(1-p)^2$	

- What will be the genotype frequencies after selfing?

- Show algebraically that the allele frequency after selfing is still p.

The obligate self-fertilization in the tiny cleistogamous flowers will result in a steady increase in the frequency of homozygous genotype but it will have no effect on the allele frequencies in the population. What happens with the showy flowers? Do their seeds fit the expectations of random mating?

Observed and predicted genotypes for chasmogamous flowers: Imagine that you collect a sample of 200 seeds that were all produced by the showy chasmogamous flowers. Fertilization of those flowers requires visits by pollinators so those seeds should represent matings between various individuals. You then determine the genotype of each seed to test whether they fit the Hardy-Weinberg expectations. Below are results for one enzyme locus, isocitrate dehydrogenase or IDH. The two alleles are labeled “1” and “2”.

IDH Genotype	Observed number	Predicted number under HWE
11	104	92.5
12	64	87.0
22	32	20.5

(Data are simplified, but are based on real information in Stewart (1994) for *I. pallida*)

What are the allele frequencies in this population?
p=(frequency of allele “1”)= _____________
q=(frequency of allele “2”)=_____________

Under HWE, we expect the proportions of the three genotypes to be p^2, 2pq and q^2. The observed allele frequencies are p=0.68 and q=0.32, so the expected genotype frequencies are p^2=0.46, 2pq=0.44, and q^2=0.10 for genotypes 11, 12, and 22 respectively. That means that in a sample of 200 individuals, we expect 92.5 “11”, 87.0 “12” and 20.5 “22” genotypes.

For this sample the observed number of seeds of each genotype does not quite match the predicted numbers. The data show an excess of the two homozygotes and a deficit of heterozygotes compared to the Hardy-Weinberg expectations. Is that difference real or could it just be a random deviation in this sample?

10.2 **Testing for departures from HWE**

If we took repeated samples of seeds from this population we would likely find slight variation in the numbers of each genotype just by chance. So it is not surprising that the numbers don’t exactly match the predicted numbers. The question is whether our results are within the normal range of variation or not. To decide whether it is unusual or not we need to do a statistical test, in this case a **chi-square (χ^2) test**. This test calculates the sum of squared deviations from the expected values and compares that sum to the value you would expect from random chance alone. The χ^2 value is calculated as followed:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed number of genotype i, and E_i is the expected number of that genotype.

IDH Genotype	**Observed number**	**Expected number**	$\frac{(O-E)^2}{E}$
11	104	92.48	1.44
12	65	87.04	5.58
22	32	20.48	6.48
Sum	**200**	**200**	**13.5**

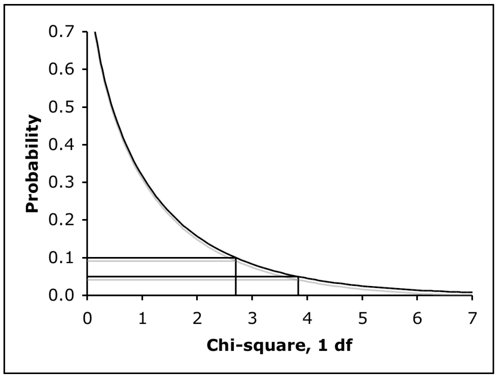
Notice that the expected values are not integers. The observed numbers will always be counts of individuals but the expected values can be, and usually are, decimal numbers. The observed and expected values will always add up to the same total number of individuals, however.

For our example,

$$\chi^2 = \frac{(104-92.48)^2}{92.48} + \frac{(65-87.04)^2}{87.04} + \frac{(32-20.48)^2}{20.48} = 13.5$$

By itself, the χ^2 value doesn't mean much. Even if the population actually fit the assumptions for HWE we would expect slight deviations in the numbers of genotypes in our particular sample of individuals just by random chance. Fortunately statisticians have calculated the probability of observing particular χ^2 values by chance alone. That probability depends on the **degrees of freedom**: the number of independent data points that we have. For this test, the degrees of freedom is the number of genotypic classes (3) minus the number of parameters estimated from our data (1, for the allele frequency p) minus 1. That leaves 1 degree of freedom.[2]

This figure shows the probability of observing a chi square value *at least as big* as this by chance alone. By convention, statisticians usually say a result is "statistically significant" if there is less than 5% of chance of it occurring by chance alone. That 5% threshold corresponds to a χ^2value of 3.84: there is only a 5% chance of observing a chi-square value at least as big as 3.84.



The observed χ^2 value of 13.5 is somewhere off the chart to the right. The probability at that point is near zero so it is extremely unlikely that the deviation from expectation is simply the result of chance. There is likely a biological reason for the deviation of genotype frequencies from the Hardy-Weinberg expectations.

Table of critical values for χ^2

DF	10%	5%	1%
1	2.706	**3.841**	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345

[2] You might wonder why we don't also have to subtract another degree of freedom for the allele frequency q. The answer is because once p is known, we also know q since it is just 1-p.

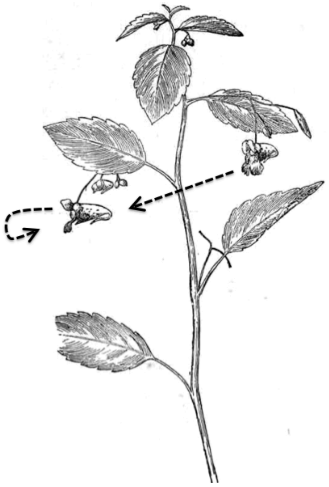
More practice: Here are similar data from the next season. Again they collected seeds from the showy chasmogamous flowers and determined the genotype of each. Calculate the allele frequencies and the expected number of each genotype. Do these data fit the Hardy Weinberg expectations?

(Remember, the chi-square test is based on the expected numbers of each genotype, not the proportions.)

IDH Genotype	Observed number	Expected number	$\frac{(O-E)^2}{E}$
11	91		
12	74		
22	35		
sum	**200**		

10.3 **Why don't the showy flowers mate randomly?**

The Chi-square test showed that there was a significant deviation from the expected genotype frequencies. In particular, there were fewer heterozygotes than expected which suggests some degree of inbreeding, even in the showy flowers. Although the showy flowers have a mechanism to prevent the transfer of pollen within a flower, it is still possible to have self-fertilization between two flowers on the same plant. If pollinators often fly short distances that kind of selfing within a plant can be common.



Yet another possibility is that there may be pollen transfer between plants that are close relatives. The seeds of this plant do not travel far so it is likely that nearby plants are offspring of the same parents. In general where there is spatial variation in allele frequencies and localized mating then there will be "inbreeding" relative to the population as a whole. That kind of population structure will also lead to an excess of homozygous genotypes.

10.4 **F_{ST} is a measure of spatial structure**

Although it would be nice to just define a population as the collection of individuals that are mating together randomly, the boundaries of a population are often very indistinct. A plant that is growing on one side of a meadow is much more likely to pollinate others growing nearby, so in that sense two sides of a meadow may be genetically separated. But when the plants are scattered evenly across that meadow, where would you draw the boundary? In addition, there may be behavioral patterns that cause individuals to mate non-randomly. If individuals live in family structured groups (e.g. wolves) they may be likely to mate with relatives.

The result is that populations are often genetically structured, at a variety of scales. Behavioral traits may result in individuals that are inbred with respect to the local subpopulation. And spatial separation may result in sub-populations that are "inbred" with respect to the population as a whole. In the latter case, inbreeding arises because individuals are on average more related to others within their own subpopulation than they are to others in more distant sites. If mating takes place within a subpopulation then they are mating with closer relatives than they would with a truly random sample of mates from the population as a whole.

In section 3.1 we used the deficit of heterozygotes to quantify inbreeding as $F = 1 - \frac{H_{obs}}{H_{\exp}}$.

Here we can use a similar approach to measure that level of inbreeding due to population structure, this time by comparing the heterozygosity within subpopulations relative to what we would expect if there were random mating over the total population.[3]

$$F_{ST} = 1 - \frac{\text{Average Expected Heterozygosity within Subpopulations}}{\text{Expected Heterozygosity of the Total Population}}$$

or

$$F_{ST} = 1 - \frac{\bar{H}_s}{H_T}$$

The heterozygosity (H) is the expected proportion of heterozygous, 2*pq*. If individuals mate at random over the entire population, then the expected total heterozygosity is simply $H_T = 2p(1-p)$.

On the other hand, if there is spatial structure and individuals mate within subpopulations then the frequency of heterozygotes will depend on the local allele frequency (p_i) in each subpopulation: $H_i = 2p_i(1-p_i)$ for subpopulation i. If there are a total of k subpopulations, then the average expected heterozygosity within subpopulations is: $\overline{H_S} = \frac{1}{k}\sum_{i=1}^{k} 2p_i(1-p_i)$.

Example: Imagine that there are two nearby patches of 200 plants each. Mating takes place only among plants from the same local area, but within those patches mating is completely random. The two subpopulations in the two areas have slightly different allele frequencies: In the first patch the frequency of allele A is p = 0.8 and in second the allele frequency is p = 0.5. But inside both sites mating is random so the genotype frequencies exactly match the expected frequencies under HWE. In the first patch the heterozygosity is 64/200 = 0.32 and in the second the heterozygosity is 100/200=0.5.

[3] Sewall Wright developed this index of population structure, which he called the "fixation index" or F. That fixation index or inbreeding coefficient can be partitioned into the various levels of population structure. F_{IS} is the fixation index of an individual with respect to its local subpopulation and F_{ST} is the average fixation index of subpopulations relative to the total population. It can be applied at any spatial scale. Sometimes people use F_{ST} to measure the differentiation among distinct populations in a large region.

Now, imagine that when we sampled the population we didn't keep track of where each plant came from. All we had was a sample of 400 plants that we assumed was from a single large population. The data we collected showed that there were a total of 178 AA homozygotes, 164 Aa heterozygotes and 58 aa homozygotes.

	Genotype			**Allele Frequency**
	AA	**Aa**	**aa**	p_A
Subpopulation 1	128	64	8	0.8
Expected	128	64	8	
Subpopulation 2	50	100	50	0.5
Expected	50	100	50	
Combined Population	178	164	58	
Expected				

What is the frequency of allele A in the combined population of all 400 plants? p =

____

What are the expected genotype frequencies in that combined sample?

____

The allele frequency in the combined population is 0.65, so the expected heterozygosity is H_T= 2*0.65*0.35 = 0.455. The average expected heterozygosity in the two subpopulations is H_S= (0.32+0.5)/2 = 0.410 so

$$F_{ST} = 1 - \frac{\bar{H}_s}{H_T} = 1 - \frac{0.410}{0.455} = 0.099$$

To understand the properties of F_{ST}, we can look at some extreme conditions. If individuals are mating completely at random over the entire population, then there will be no local variation in allele frequency and each of the subpopulations will have the same expected heterozygosity as the total population. In that case Fst will be 0: there is no differentiation among subpopulations. At the other extreme, each subpopulation may be completely isolated from all of the other subpopulations and each subpopulation may have become fixed for a different allele. That is the maximum level of differentiation. If each subpopulation is fixed for one allele or the other, then there is no heterozygosity within subpopulations (H_S =0). In that case Fst will be 1.0.

Real populations are never that extreme. For most species of animals, F_{ST} ranges from 0 to 0.2. Plant populations usually show somewhat higher degrees of spatial structure because plants are rooted in place. Yet even for plants F_{ST} is usually less than 0.4.

But biologists can use Fst as a measure of population structure and to identify distinct subpopulations. For example, you saw in a previous chapter how salmon populations can be identified by differences in allele frequency. That variation can also be used to measure the degree of connectedness among populations. In one study of Fraser River populations of

Chinook salmon the average Fst within the upper or lower sections of the river was low (only 0.017 and 0.001, respectively). That suggests that the subpopulations were not very distinct within each region. However the Fst value between the upper vs lower Fraser River regions was much higher (0.069). Biologists used that to support the hypothesis that the upper vs lower river populations should be considered as genetically distinct regions.

10.5 **Your turn:**

Researchers measured the allele frequencies in several patches of jewelweed to quantify the degree of local population structure.

Here are a subset of their results:

Site	allele frequency p_i	Expected Heterozygosity H_i
1	0.4	
2	0.2	
3	0.35	
4	0.7	
5	0.6	

The combined allele frequency for the entire area was 0.45. What is the value of F_{ST}?

Answers

p 3. After selfing, the genotype frequencies will be: $p^2+\frac{1}{2}p(1-p)$, $p(1-p)$, and $(1-p)^2+\frac{1}{2}p(1-p)$.

Calculate the allele frequency as $p = P + \frac{1}{2}H$ so

$$
\begin{aligned}
p' &= p^2 + \frac{1}{2}p(1-p) + \frac{1}{2}p(1-p) \\
&= p^2 + p(1-p) \\
&= p^2 + p - p^2 \\
&= p
\end{aligned}
$$

p 4. p=(104+(64/2)) / 200 = 0.68; q= (32 + (64/2)) / 200=0.32

p 6.

IDH Genotype	Observed number	Expected number	$\frac{(O-E)^2}{E}$
11	91	81.92	1.01
12	74	92.16	3.58
22	35	25.92	3.18
sum	**200**	**200**	**7.77**

p 8. The allele frequency in the combined population is p=0.65

The expected numbers for each genotype are 169, 82, and 49

p 9.

Site	allele frequency p_i	Expected Heterozygosity H_i
1	0.4	0.48
2	0.2	0.32
3	0.35	0.455
4	0.7	0.42
5	0.6	0.48

H_S = 0.431. H_T = 0.495. F_{ST} = 1 - 0.431/0.495 = 0.129