Population Structure

Timothy Thornton

Summer Institute in Statistical Genetics 2013
Module 8
Lecture 9

Nonrandom Mating

- HWE assumes that mating is random in the population
- Most natural populations deviate in some way from random mating
- There are various ways in which a species might deviate from random mating
- We will focus on the two most common departures from random mating:
  - inbreeding
  - population subdivision or substructure

Nonrandom Mating: Inbreeding

- Inbreeding occurs when individuals are more likely to mate with relatives than with randomly chosen individuals in the population
- Increases the probability that offspring are homozygous, and as a result the number of homozygous individuals at genetic markers in a population is increased
- Increase in homozygosity can lead to lower fitness in some species
- Increase in homozygosity can have a detrimental effect: For some species the decrease in fitness is dramatic with complete infertility or inviability after only a few generations of brother-sister mating

Nonrandom Mating: Population Subdivision

- For subdivided populations, individuals will appear to be inbred due to more homozygotes than expected under the assumption of random mating.
- Wahlund Effect: Reduction in observed heterozygosity (increased homozygosity) because of pooling discrete subpopulations with different allele frequencies that do not interbreed as a single randomly mating unit.

Wright's F Statistics

- Sewall Wright invented a set of measures called F statistics for departures from HWE for subdivided populations.
- F stands for fixation index, where fixation being increased homozygosity
- F_{IS} is also known as the inbreeding coefficient.
  - The correlation of uniting gametes relative to gametes drawn at random from within a subpopulation (**I**ndividual within the **S**ubpopulation)
- F_{ST} is a measure of population substructure and is most useful for examining the overall genetic divergence among subpopulations
  - Is defined as the correlation of gametes within subpopulations relative to gametes drawn at random from the entire population (**S**ubpopulation within the **T**otal population).

Wright's F Statistics

- F_{IT} is not often used. It is the overall inbreeding coefficient of an individual relative to the total population (**I**ndividual within the **T**otal population).

Genotype Frequencies for Inbred Individuals

- Consider a bi-allelic genetic marker with alleles A and a. Let p be the frequency of allele A and $q = 1 - p$ the frequency of allele a in the population.
- Consider an individual with inbreeding coefficient F. What are the genotype frequencies for this individual at the marker?

Genotype	AA	Aa	aa
Frequency			

Generalized Hardy-Weinberg Deviations

- The table below gives genotype frequencies at a marker for when the HWE assumption does not hold:

Genotype	AA	Aa	aa
Frequency	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$

  where $q = 1 - p$
- The F parameter describes the deviation of the genotype frequencies from the HWE frequencies.
- When $F = 0$, the genotype frequencies are in HWE.
- The parameters p and F are sufficient to describe genotype frequencies at a single locus with two alleles.

F_{st} for Subpopulations

- Example in Gillespie (2004)
- Consider a population with two equal sized subpopulations. Assume that there is random mating within each subpoulation.
- Let $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$
- Below is a table with genotype frequencies

Genotype	A	AA	Aa	aa
Freq. Subpop$_1$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop$_2$	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$

- Are the subpopulations in HWE?
- What are the genotype frequencies for the entire population?
- What should the genotypic frequencies be if the population is in HWE at the marker?

F_{st} for Subpopulations

- Fill in the table below. Are there too many homozygotes in this population?

	Allele	Genotype		
	A	AA	Aa	aa
Freq. Subpop_1	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop_2	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$
Freq. Population				
Hardy-Weinberg Frequencies				

- To obtain a measure of the excess in homozygosity from what we would expect under HWE, solve

$$2pq(1 - F_{ST}) = \frac{3}{8}$$

- What is F_{st}?

F_{st} for Subpopulations

- The excess homozygosity requires that $F_{ST} =$ _____
- For the previous example the allele frequency distribution for the two subpopulations is given.
- At the population level, it is often difficult to determine whether excess homozygosity in a population is due to inbreeding, to subpopulations, or other causes.
- European populations with relatively subtle population structure typically have an F_{st} value around .01 (e.g., ancestry from northwest and southeast Europe),
- F_{st} values that range from 0.1 to 0.3 have been observed for the most divergent populations (Cavalli-Sforza et al. 1994).

F_{st} for Subpopulations

- Nelis et al. (PLOS One, 2009) looked at the genetic structure for various populations
- Obtained pairwise F_{st} values for the four HapMap sample populations
  - Europeans (CEU) - Africans (YRI): **0.153**
  - Europeans (CEU) - Japanese (JPT): **0.111**
  - Europeans (CEU) - Chinese (CHB): **0.110**
  - Africans (YRI) - Chinese (CHB): **0.190**
  - Africans (YRI) - Japanese (JPT): **0.192**
  - Chinese (CHB) - Japanese (JPT): **0.007**

F_{st} for Subpopulations

- F_{st} can be generalized to populations with an arbitrary number of subpopulations.
- The idea is to find an expression for F_{st} in terms of the allele frequencies in the subpopulations and the relative sizes of the subpopulations.
- Consider a single population and let r be the number of subpopulations.
- Let p be the frequency of the A allele in the population, and let p_i be the frequency of A in subpopulation i, where $i = 1, \ldots, r$
- F_{st} is often defined as $F_{st} = \frac{\sigma_p^2}{p(1-p)}$, where σ_p^2 is the variance of the p_i's with $E(p_i) = p$.

F_{st} for Subpopulations

- Let the relative contribution of subpopulation i be c_i, where $\sum_{i=1}^{r} c_i = 1$.

Genotype	AA	Aa	aa
Freq. Subpop$_i$	p_i^2	$2p_iq_i$	q_i^2
Freq. Population	$\sum_{i=1}^{r} c_i p_i^2$	$\sum_{i=1}^{r} c_i 2p_iq_i$	$\sum_{i=1}^{r} c_i q_i^2$

  where $q_i = 1 - p_i$
- In the population, we want to find the value F_{st} such that $2pq(1 - F_{st}) = \sum_{i=1}^{r} c_i 2p_iq_i$
- Rearranging terms:

$$F_{st} = \frac{2pq - \sum_{i=1}^{r} c_i 2p_iq_i}{2pq}$$

- Now $2pq = 1 - p^2 - q^2$ and $\sum_{i=1}^{r} c_i 2p_iq_i = 1 - \sum_{i=1}^{r} c_i(p_i^2 + q_i^2)$

F_{st} for Subpopulations

- So can show that

$$F_{st} = \frac{\sum_{i=1}^{r} c_i(p_i^2 + q_i^2) - p^2 - q^2}{2pq}$$

$$= \frac{\left[\sum_{i=1}^{r} c_i p_i^2 - p^2\right] + \left[\sum_{i=1}^{r} c_i q_i^2 - q^2\right]}{2pq}$$

$$= \frac{Var(p_i) + Var(q_i)}{2pq}$$

$$= \frac{2Var(p_i)}{2p(1-p)}$$

$$= \frac{Var(p_i)}{p(1-p)}$$

$$= \frac{\sigma_p^2}{p(1-p)}$$

Estimating F_{st}

- Let n be the total number of sampled individuals from the population and let n_i be the number of sampled individuals from subpopulation i
- Let $\hat{p}_i$ be the allele frequency estimate of the A allele for the sample from subpopulation i
- Let $\hat{p} = \sum_i \frac{n_i}{n} \hat{p}_i$
- A simple F_{st} estimate is $\hat{F}_{ST_1} = \frac{s^2}{\hat{p}(1-\hat{p})}$, where s^2 is the sample variance of the $\hat{p}_i$'s.

Estimating F_{st}

- Weir and Cockerman (1984) developed an estimate based on the method of moments.

$$MSA = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\hat{p}_i - \hat{p})^2$$

$$MSW = \frac{1}{\sum_i (n_i - 1)}\sum_{i=1}^{r} n_i \hat{p}_i(1 - \hat{p}_i)$$

- Their estimate is

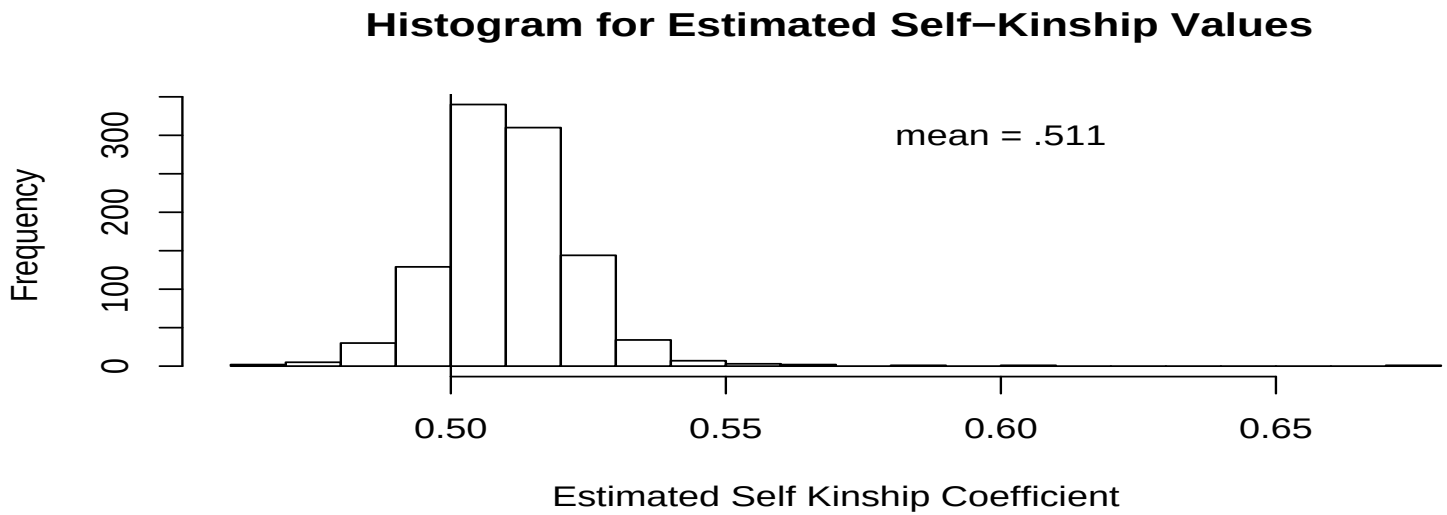
$$\hat{F}_{ST_2} = \frac{MSA - MSW}{MSA + (n_c - 1)MSW}$$

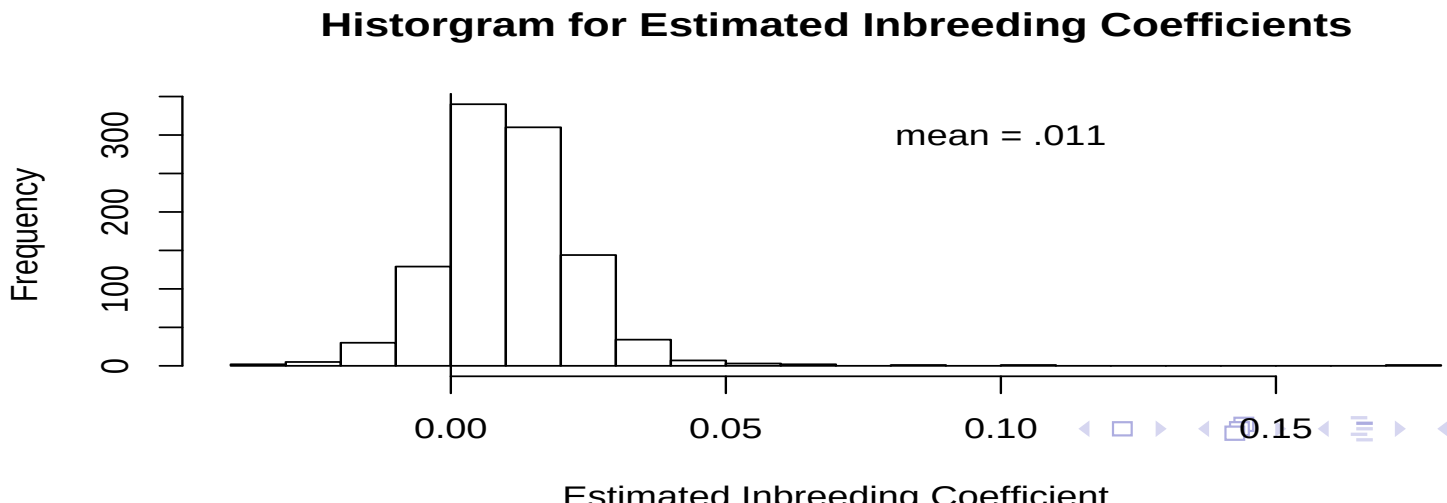
where $n_c = \sum_i n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$

GAW 14 COGA Data

- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 1,009 individuals from 143 pedigrees with each pedigree containing at least 3 affected individuals.
- Individuals labeled as white, non-Hispanic were considered.
- Estimated self-kinship and inbreeding coefficients using genome-screen data

COGA Data


Histogram for Estimated Self−Kinship Values
mean = .511
Frequency
0
100
200
300
0.50
0.55
0.60
0.65
Estimated Self Kinship Coefficient


Historgram for Estimated Inbreeding Coefficients
mean = .011
Frequency
0
100
200
300
0.00
0.05
0.10
0.15
Estimated Inbreeding Coefficient

References

- Nelis M, Esko T, Magi R, Zimprich F, Zimprich A, et al. (2009) Genetic Structure of Europeans: A View from the NorthEast. *PLoS ONE* **4**, e5472. doi:10.1371/journal.pone.0005472.
- Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.