

INTRODUCTION TO *F* (OR *G*) STATISTICS

READING: Nielsen & Slatkin pp. 60-63

- *F* and *G* statistics are used to
 - describe levels of genetic diversity that occur within and between subpopulations
 - estimate levels of gene flow

- *F* and *G* stats are often called **diversity Indices** since they are based on actual, potential, and/or virtual heterozygosities

- Imagine a collection of subpopulations

- Let $p_{s,j}$ = frequency of allele A_j in subpopulation s . $j = 1, \dots, k$ (i.e., there are k alleles)
 H_s = observed frequency of heterozygotes in subpopulation s ;

- Define the following

(1) $H_I = \text{avg}_s(H_s)$ = average frequency of heterozygotes over subpopulations.

(2) $H_{s,s}$ = Expected frequency of "heterozygotes" in subpopulation s assuming random union of gametes

$$= 1 - \sum_{j=1}^k p_{j,s}^2$$

(3) $\bar{H}_s = \text{avg}_s(H_{s,s}) = 1 - \sum_{j=1}^k \bar{p}_j^2$

where $\bar{p}_j^2 = \text{avg}_s(p_{j,s}^2)$ = Average frequency of homozygotes (of any kind) expected under R.U.G.

(4) H_T = expected frequency of heterozygotes if entire population were to mate at random

$$= 1 - \sum_{j=1}^k (\bar{p}_j)^2 \quad \text{where } \bar{p}_j = \text{avg}_s(p_{j,s}).$$

Observation 1: \bar{H}_s and H_T can be computed for haploids and polyploids, in which case they represent *virtual* (instead of actual) expected heterozygosities.

Observation 2: With two alleles, $H_{s,s} = 2p_sq_s$ and $H_T = 2\bar{p}\bar{q}$

Comment: When *estimating* these diversity indices, these formulae are not statistically optimal since they don't account for sampling error.

– Now, let's define the F statistics (F_{IS} , F_{ST} , F_{IT}) themselves...

$$(1) \quad F_{IS} = \frac{\bar{H}_s - H_I}{\bar{H}_s} \quad = \text{reduction in heterozygosity due to nonrandom mating between relatives within subpopulations}$$

$$(2) \quad F_{ST} = \frac{H_T - \bar{H}_s}{H_T} \quad = \text{reduction in heterozygosity due to population subdivision (the Wahlund effect, as we'll see)}$$

$$(3) \quad F_{IT} = \frac{H_T - H_I}{H_T} \quad = \text{reduction in heterozygosity due to nonrandom mating between relatives within subpopulations and population subdivision (again, Wahlund effect)}$$

- "Reduction in heterozygosity" \equiv "Inbreeding"
- Only F_{ST} can be computed for non-diploids
- When there are > 2 alleles, F_{ST} is called G_{ST} .
- Again, different formulae are used in practice to actually *estimate* F statistics

EXAMPLE: "Wahlund Effect"

- Moral: F_{ST} measures reduction in heterozygosity due to population subdivision
- Recall:
 - we have a collection of infinitely-sized subpopulations
 - two alleles, A and a with frequencies $p_{1,s}$ ($= "p"$) and $p_{2,s}$ ($= "q"$) in subpopulation s
 - within subpopulations, there is random mating
- Then, H_I = the average observed frequency of heterozygotes in subpopulation s =

$$\text{avg}_s(2p_{1,s}p_{2,s}) = "2\bar{p}\bar{q} - 2\text{Var}(p)" \text{ (as we saw several weeks ago) = } \\ 2\bar{p}_1\bar{p}_2 - 2\text{Var}(p_1) \text{ (using the new lingo)}$$
- $H_{s,s}$ = expected freq. of hets. under R.U.G = $2p_{1,s}p_{2,s}$

- $\bar{H}_s = \text{avg}_s(H_{s,s}) = \text{avg}_s(2p_{1,s}p_{2,s}) = 2\bar{p}_1\bar{p}_2 - 2\text{var}(p_1) = H_I!$

- Finally, $H_T = 2\bar{p}_1\bar{p}_2$

- So,
$$F_{IS} = \frac{H_s - H_I}{H_s} = \frac{[2\bar{p}_1\bar{p}_2 - 2\text{var}(p_1)] - [2\bar{p}_1\bar{p}_2 - 2\text{var}(p_1)]}{2\bar{p}_1\bar{p}_2 - 2\text{var}(p_1)} = 0.$$

$$F_{ST} = \frac{[2\bar{p}_1\bar{p}_2] - [2\bar{p}_1\bar{p}_2 - 2\text{var}(p_1)]}{2\bar{p}_1\bar{p}_2} = \frac{\text{var}(p_1)}{\bar{p}_1\bar{p}_2}.$$

Also,
$$F_{IT} = \frac{2\bar{p}_1\bar{p}_2 - [2\bar{p}_1\bar{p}_2 - 2\text{var}(p_1)]}{2\bar{p}_1\bar{p}_2} = \frac{\text{var}(p_1)}{\bar{p}_1\bar{p}_2} = F_{ST}$$

– **Conclusions:**

- All inbreeding is due to population subdivision, none due to nonrandom mating between relatives; (i.e., $F_{IT} = F_{ST}$)
- F_{ST} is the inbreeding coefficient, f , we computed previously for the Wahlund effect.

– Typical values of F_{ST}