Chapter 5

Population Structure

Introduction

(1) Statistical and genetic sampling :

Population genetic theory depends on the concept of replicated populations that are maintained under the same conditions, but will differ because of genetic sampling. The derived variance for statistics of interest should consider both types of variation.

(2) Between-population variation cannot be estimated with a sample from a single population. Different loci sometimes may be regarded as playing the role of separate populations.

(3) **The distinction between statistical and genetic sampling can also be phrased in terms of fixed and random effects.**

Fixed populations

Every member of the population has an equal chance of being sampled, and individuals are sampled independently, the genotypic counts are multinomially distributed.

Different populations for the same species are compared simply by comparing frequencies.

Contingency Tables: With v alleles at a locus, the genotypic counts in each of r samples are arranged in a $v(v+1)/2 \times r$ contingency table and a chi-square statistic with $[v(v+1)/2-1]\times r$ degrees of freedom is calculated.

Numerical Resampling:

Bootstrapping: Two populations can be judged to have different allele frequencies if the estimated frequencies have nonoverlapping confidence intervals.

Chebyshev's inequality:

$$\Pr(\left|\tilde{p}_{i_I} - p_{i_I}\right| > k\sigma) \leq \frac{1}{k^2}$$

$$\Pr(\left|\tilde{p}_{i_I} - p_{i_{II}}\right| > K\sigma\sqrt{2}) \leq \frac{1}{K^2}$$

Permutation Tests:

$$\Pr(\{n_{i_I}\},\{n_{i_{II}}\}) = \left(\frac{(2n_I)!}{\prod_i n_{i_I}!}\prod_i p_{iI}^{n_{iI}}\right)\left(\frac{(2n_{II})!}{\prod_i n_{i_{II}}!}\prod_i p_{iII}^{n_{iII}}\right)$$

$$\Pr(\{n_{i_I}\},\{n_{i_{II}}\}\,|\,\{n_{i_I}+n_{i_{II}}\}) = \frac{(2n_I)!(2n_{II})!}{(2n_I+2n_{II})!}\frac{\prod_i (n_{i_I}+n_{i_{II}})!}{\prod_i n_{i_I}!\prod_i n_{i_{II}}!}$$

	Sample I	Sample II
Allele A	$n_{A_I}=8$	$n_{A_{II}}=2$
Allele a	$n_{a_I}=2$	$n_{a_{II}}=8$
	$2n_I=10$	$2n_{II}=10$

Table 5.1 Possible allocations of 10 A and 10 a alleles into two samples of size 10 alleles, along with conditional probabilities.

Number of A alleles		
Sample I	Sample II	Probability
0	10	0.0000
1	9	0.0005
2	8	0.0110
3	7	0.0779
4	6	0.2387
5	5	0.3438
6	4	0.2387
7	3	0.0779
8	2	0.0110
9	1	0.0005
10	0	0.0000

F statistics:

$$F_{st} = \frac{\sum_i (\tilde{P}_i - \overline{P})^2 \Big/ (r-1)}{\overline{P}(1-\overline{P})}$$

In the fixed-population framework,

$$X^2 = (r-1)\overline{n}F_{st}$$

Analysis of Variance:

$$X_{ij} = \begin{cases} 1 & \text{if allele is A} \\ 0 & \text{if allele is not A} \end{cases}$$

Table 5.2 Analysis of variance layout for variable indicating allele A in fixed populations.

Source	d.f.	Sum of Squares	Expected Mean Square*
Between populations	$r-1$	$\sum_i \frac{x_{i\cdot}^2}{n_i} - \frac{x_{\cdot\cdot}^2}{n_\cdot}$	$\frac{1}{r-1}\sum_i k_1 p_{Ai}(1-p_{Ai})$
		$= \sum_i n_i(\tilde{p}_{Ai} - \tilde{p}_{A\cdot})^2$	$+\frac{1}{r-1}\sum_i n_i(p_{Ai} - \bar{p}_{A\cdot})^2$
Within populations	$\sum_i (n_i - 1)$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}^2 - \sum_i \frac{x_{i\cdot}^2}{n_i}$	$\frac{1}{\sum_i (n_i-1)}\sum_i k_2 p_{Ai}(1-p_{Ai})$
	$= n_\cdot - r$	$= \sum_i n_i \tilde{p}_{Ai}(1-\tilde{p}_{Ai})$	

*$k_1 = \left(1 - \frac{n_i}{\sum_i n_i}\right)$, $k_2 = (n_i - 1)$

Random Populations

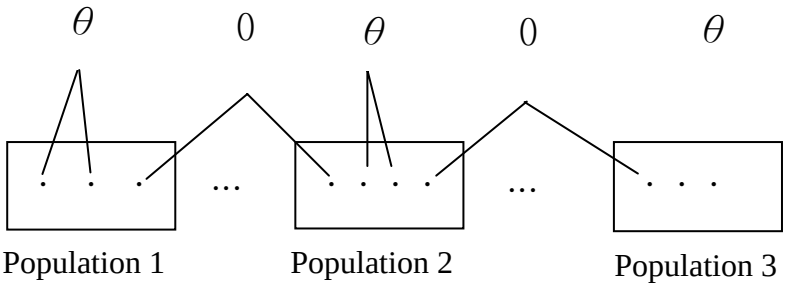
The action of evolutionary forces (genetic sampling) causes different alleles in a population to be dependent, or related, and will result in intraspecific differentiation. Even though individuals, or alleles, may be sampled randomly, the process of taking expectations must recognize that they are dependent through their shared ancestry.

The differentiation is quantified with the F statistics of Wright (1951), or the analogous measures of Cockerham (1963,1973).

$F(F_{IT})$

$\theta(F_{ST})$

$f(F_{IS})$



Haploid Data

There is only one F statistic: θ

Shared ancestry means that the expected value of a squared sample frequency from a sample of size n_i is

$$E(\tilde{P}_{Ai}^2) = P_A^2 + P_A(1-P_A)\theta + \frac{1}{n_i}P_A(1-P_A)(1-\theta)$$

Between population differentiation goes hand in hand with relatedness of alleles within populations. As individuals become more related within populations, the independent populations are expected to become more differentiated.

Table 5.3 Analysis of variance layout for variable indicating allele A in random populations.

Source	d.f.	Sum of Squares	Expected Mean Square*
Between populations	$r-1$	$\sum_{i=1}^{r}\frac{x_{i\cdot}^2}{n_i}-\frac{x_{\cdot}^2}{n_{\cdot}}$	$p_A(1-p_A)[(1-\theta)+n_c\theta]$
		$=\sum_{i=1}^{r} n_i(\tilde{p}_{Ai}-\tilde{p}_{A\cdot})^2$	$=\sigma_G^2+n_c\sigma_P^2$
Within populations	$\sum_{i=1}^{r}(n_i-1)$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}^2-\sum_i\frac{x_{i\cdot}^2}{n_i}$	$p_A(1-p_A)(1-\theta)$
	$=n_{\cdot}-r$	$=\sum_i n_i\tilde{p}_{Ai}(1-\tilde{p}_{Ai})$	$=\sigma_G^2$

*$n_c=\frac{1}{r-1}\left(\sum_{i=1}^{r}n_i-\frac{\sum_i n_i^2}{\sum_i n_i}\right)$

$$p_A(1-p_A)(1-\theta) \;\hat{=}\; \frac{1}{\sum_i n_i - 1}\sum_i n_i\tilde{p}_{Ai}(1-\tilde{p}_{Ai}) \;=\; \text{MSG}$$

$$p_A(1-p_A)(1-\theta)+n_c p_A(1-p_A)\theta \;\hat{=}\; \frac{1}{r-1}\sum_i n_i(\tilde{p}_{Ai}-\tilde{p}_{A\cdot}) \;=\; \text{MSP}$$

$$\sigma_G^2 \;=\; p_A(1-p_A)(1-\theta) \qquad \sigma_P^2 \;=\; p_A(1-p_A)\theta$$

$$\hat{\sigma}_G^2 \;=\; \text{MSG} \qquad \hat{\sigma}_P^2 \;=\frac{1}{\mathrm{n_c}}\text{(MSP-MSG)}$$

An estimate of θ from ANOVA is

$$\hat{\theta} \;=\; \frac{\hat{\sigma}_P^2}{\hat{\sigma}_P^2+\hat{\sigma}_G^2} \;=\; \frac{\text{MSP - MSG}}{MSP+(n_c-1)MSG}$$

For a large sample,

$$\hat{\theta} \;=\; \frac{S_A^2}{\tilde{P}_{A\bullet}(1-\tilde{P}_{A\bullet})} \;=\; F_{ST}$$

For multiple loci and alleles, see pg174.

Some comments on θ (pg 174 ~ 176)

Diploid Data

Table 5.4 Analysis of variance layout for genotypic data in random populations.

Source	d.f.	Sum of Squares	Expected Mean Square*
Between populations	$r-1$	$2\sum_{i=1}^{r} n_i(\tilde{p}_{Ai}-\tilde{p}_{A\cdot})^2$	$p_A(1-p_A)\,[(1-F)$
		$= 2(r-1)\bar{n}s_A^2$	$+2(F-\theta)+2n_c\theta]$
			$=\sigma_G^2+2\sigma_I^2+2n_c\sigma_P^2$
Individuals in populations	$\sum_{i=1}^{r}(n_i-1)$	$\sum_{i=1}^{r} n_i(\tilde{p}_{Ai}+\tilde{P}_{AAi}-2\tilde{p}_{Ai}^2)$	$p_A(1-p_A)\,[(1-F)$
	$= n_\cdot - r$	$= 2r\bar{n}\tilde{p}_{A\cdot}(1-\tilde{p}_{A\cdot})-\frac{1}{2}r\bar{n}\tilde{H}_{A\cdot}$	$+2(F-\theta)]$
		$-2(r-1)\bar{n}s_A^2$	$=\sigma_G^2+2\sigma_I^2$
Alleles in individuals	$\sum_{i=1}^{r} n_i$	$\sum_{i=1}^{r} n_i(\tilde{p}_{Ai}-\tilde{P}_{AAi})$	$p_A(1-p_A)(1-F)$
	$= n_\cdot$	$=\frac{1}{2}r\bar{n}\tilde{H}_{A\cdot}$	$=\sigma_G^2$

*$n_c = \frac{1}{r-1}\left(\sum_{i=1}^{r} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}\right)$

$$E(\tilde{P}_{Ai}^2) = P_A^2 + P_A(1-P_A)\theta + \frac{1}{2n_i}P_A(1-P_A)(1-F-2\theta)$$

$\sigma_P^2 = P_A(1-P_A)\theta$ for populations

$\sigma_I^2 = P_A(1-P_A)(F-\theta)$ for individuals within populations

$\sigma_G^2 = P_A(1-P_A)(1-F)$ for alleles within individuals

$\hat{\sigma}_G^2 =$ MSG $\quad \hat{\sigma}_I^2 = \frac{1}{2}$(MSI - MSP)

$\hat{\sigma}_P^2 = \frac{1}{2n_c}$(MSP - MSI)

Estimation of F, θ , f (pg177~179)

$$\hat{F} = \frac{\hat{\sigma}_P^2 + \hat{\sigma}_I^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2}$$

$$= 1 - \frac{2n_c \text{MSG}}{\text{MSP} + (n_c - 1)\text{MSI} + n_c \text{MSG}}$$

$$= 1 - \frac{S_3}{S_2}$$

$$\hat{F} = \frac{\hat{\sigma}_P^2}{\hat{\sigma}_P^2 + \hat{\sigma}_I^2 + \hat{\sigma}_G^2}$$

$$= \frac{\text{MSP} - \text{MSI}}{\text{MSP} + (n_c - 1)\text{MSI} + n_c \text{MSG}}$$

$$= \frac{S_1}{S_2}$$

$$\hat{F} = 1 - \frac{\tilde{H}_A}{\tilde{p}_{A\cdot}(1 - \tilde{p}_{A\cdot})} \qquad \hat{\theta} = \frac{s_A^2}{\tilde{p}_{A\cdot}(1 - \tilde{p}_{A\cdot})}$$

$$\hat{f} = \frac{\hat{F} - \hat{\theta}}{1 - \hat{\theta}}$$

Effects of Evolutionary Forces (on θ)

When different mating system and mutation or migration process are specified, the expected

variance of θ can be estimated by the methods of Cockerham and Weir (1983)

Under random mating, F= θ and $f = 0$.

Any avoidance of mating between relatives will cause F< θ and f<0.

Different patterns of differences for the two estimates of F and θ at different loci indicate that there are forces other than nonrandom mating affecting these loci.

The effects of selection on the F statistics were detailed by Cockerham (1973).

If forces such as mutation are involved,

….(pg 180 ~183)

$$\theta = \frac{1}{1+4N\mu}$$

Migration : $\theta = \frac{1}{1+4Nm}$

Population Subdivision

Table 5.5 Analysis of variance layout for a three-level sampling hierarchy in random populations.

Source	d.f.	M.S.	Expected M.S.*
Between populations	$r-1$	MSP	$(1-F)+2(F-\theta_S)$ $+2n_{c1}(\theta_S-\theta_P)+2n_{c2}\theta_P$
Subpopulations in populations	$\sum_{i=1}^{r}(s_i-1)$ $=s.-r$	MSS	$(1-F)+2(F-\theta_S)$ $+2n_{c3}(\theta_S-\theta_P)$
Individuals in subpopulations	$\sum_{i=1}^{r}\sum_{j=1}^{s_i}(n_{ij}-1)$ $=n..-s.$	MSI	$(1-F)+2(F-\theta_S)$
Alleles in individuals	$\sum_{i=1}^{r}\sum_{j=1}^{s_i}n_{ij}$ $=n..$	MSG	$(1-F)$

*To be multiplied by $p_A(1-p_A)$. n_{c1}, n_{c2}, n_{c3} are defined in the text.

Three-Level Hierarchy
Four-Level Hierarchy

Genetic Distance

Geometric distance and Genetic distance

Geometric Distances :

$$d_{P\theta} = P\sqrt{\sum_{u=1}^{v}(P_u - q_u)^2}$$

Coancestry as Distance

$$\theta_{t+1} = \frac{1}{2N} + (1-\frac{1}{2N})\theta_t$$

If $\theta_0 = 0$, $\theta_t = 1-(1-\frac{1}{2N})^t$

will give information about *t*, the time since the populations diverged.

Specifically

$$d = -\ln(1-\theta) = -t\ln(1-\frac{1}{2N}) \approx \frac{t}{2N}$$

is an appropriate distance for divergence due to drift.

For equal sample sizes from two populations in which the frequencies of allele A_u at locus l, the estimator becomes

$$\hat{\theta} = \frac{\sum_l \left\{ \frac{1}{2}\sum_u (\tilde{p}_{lu_1} - \tilde{p}_{lu_2})^2 - \frac{1}{2(2n-1)}\left[2 - \sum_u (\tilde{p}^2_{lu_1} + \tilde{p}^2_{lu_2}) \right] \right\}}{\sum_l (1 - \sum_u \tilde{p}_{lu_1}\tilde{p}_{lu_2})}$$

Nei's Genetic Distance

$$I = \frac{\sum_l \sum_u \tilde{P}^2_{lu_1} - \tilde{P}^2_{lu_2}}{\sqrt{\sum_l \sum_u \tilde{P}^2_{lu_1} \sum_l \sum_u \tilde{P}^2_{lu_2}}}$$

Nei's distance is appropriate for long-term evolution when populations diverge because of drift and mutation. The distance is proportional to the time since divergence in the special case of the infinite alleles mutation model and equilibrium in the ancestry population. The Coancestry distance is appropriate divergence due to drift only, and no assumptions need to be made about the ancestral population.

Variance of Distance Estimates

Summary

Homework: Exercise 5.3 (pg 200)