Chapter 2 Estimating Frequencies

Estimation

estimator, consistency, unbias, variance, efficiency and sufficiency

Multinomial Genotypic Counts

Multinomial distribution & hypergeometric distribution

$$\Pr(n_1, n_2, \cdots, n_k) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} \theta_i^{n_i}$$

Multinomial moments

$$E(n_i) = \sum_{r=0}^{n} rP_r(n_i = r) = n\theta_i$$

$$E(\tilde{\theta}_i) = \theta_i$$

$$Var(n_i) = n\theta_i(1-\theta_i)$$

$$Var(\tilde{\theta}_i) = \frac{1}{n}\theta_i(1-\theta_i)$$

$$E(n_i n_j) = n(n-1)\theta_i\theta_j$$

$$E(\tilde{\theta}_i\tilde{\theta}_j) = \frac{(n-1)}{n}\theta_i\theta_j$$

$$Cov(n_i, n_j) = -n\theta_i\theta_j$$

$$Cov(\tilde{\theta}_i, \tilde{\theta}_j) = -\frac{1}{n}\theta_i\theta_j$$

$$Corr\ (n_i, n_j) = -\frac{\theta_i \theta_j}{\sqrt{\theta_i (1-\theta_i)\theta_j (1-\theta_j)}}$$

$$= Corr\ (\tilde{\theta}_i, \tilde{\theta}_j)$$

Within-population variance of allele frequencies

$$n_u = 2n_{uu} + \sum_{u \neq v} n_{uv}$$

$$E\ (n_u) = 2\ nP_{uu} + \sum_{u \neq v} nP_{uv} = 2\ nPu$$

Where $P_u = P_{uu} + \sum_{u \neq v} P_{uv}$

$$Var\ (n_u) = 2n(P_u + P_{uu} - 2P_u^2)$$

$$Var\ (\tilde{P}_u) = \frac{1}{2n}(P_u + P_{uu} - 2P_u^2)$$

Confidence interval for $\tilde{P}_u$

Under Hardy-Weinberg equilibrium

$$P_{uu} = P_u^2 \qquad P_{uv} = 2P_u P_v$$

$$Var\ (\tilde{P}_u) = \frac{1}{2n} P_u (1 - P_u)$$

Indicator variables

$$x_{ij} = \begin{cases} 1, \text{ if allele } j \text{ of individual } i \text{ B type A} \\ 0, \text{ otherwise} \end{cases}$$

$$\tilde{P}_A = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} x_{ij}$$

$$E(x_{ij}) = 1 \times \Pr(x_{ij} = 1) + 0 \times \Pr(x_{ij} = 0)$$

$$= 1 \times P_A + 0 \cdot (1 - P_A) = P_A$$

$$E(\tilde{P}_A) = P_A$$

$$E(x_{ij}^2) = P_A$$

$$E(x_{ij}x_{i\,j'}) = P_{AA}$$

$$E(x_{ij}x_{i'j'}) = E(x_{ij})E(x_{i'j'}) = P_A^2$$

$$E(\tilde{P}_A^2) = P_A^2 + \frac{1}{2n}(P_A + P_{AA} - 2P_A^2)$$

$$Var(\tilde{P}_A) = \frac{1}{2n}(P_A + P_{AA} - 2P_A^2)$$

Within-population covariance of allele frequencies

$$Cov(\tilde{P}_1, \tilde{P}_2) = -\frac{1}{2n}P_1P_2$$

Example:

(1) MN blood groups
(2) Pgm locus in mosquito
(3) esterase locus

Total variance of allele frequencies

Genetic sampling & statistical sampling
Dirichlet & multinomial distribution

$$E(x_{ij}x_{i'j'}) = P_{A/A}$$

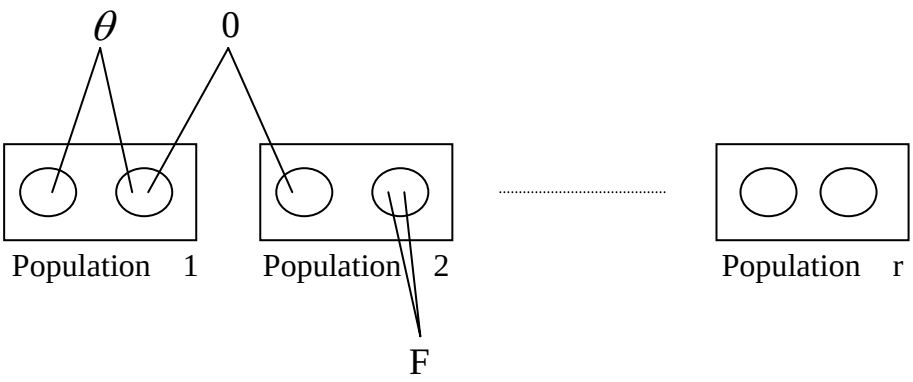
$$Var(\tilde{P}_A) = (P_{A/A} - P_A^2) + \frac{1}{2n}(P_A + P_{AA} - 2P_{A/A})$$

$$P_{AA} = P_A^2 + P_A(1 - P_A)F$$

$$P_{A/A} = P_A^2 + P_A(1 - P_A)\theta$$

$$Var(\tilde{P}_A) = P_A(1-P_A)\left(\theta + \frac{F-\theta}{n} + \frac{1-F}{2n}\right)$$

group coancestry coefficient θ_L



In a random mating population

$$Var(\tilde{P}_A) = P_A(1-P_A)\theta + P_A(1-P_A)\frac{1-\theta}{2n}$$

$$= P_A(1-P_A)\theta$$

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

With random mating, $F = \theta$ and $f = 0$

Fisher's approximate variance formula

Numerical resampling: The Jackknife and Bootstrap

Maximum likelihood estimation

$L(P_A)$ & $L(P_A, f)$

Properties of MLE

Bailey's method for MLE

$$\hat{f} = 1 - \frac{n_{Aa}}{2n\hat{P}_A(1-\hat{P}_A)}$$

$$Var(\hat{f}) = \frac{1}{n}(1-f)^2(1-2f) + \frac{f(1-f)(2-f)}{2nP_A(1-P_A)}$$

Iterative solutions of likelihood equations
The EM algorithm

Example 1: Frequencies of genotypes and phenotypes for ABO blood groups. Alleles A,B,O have frequencies p,q and r.

Genotype	Phenotype	Count	Expected Frequency	Estimated Count
AA	A	n_A	p^2	n^*_{AA}=[p/(p+2r)] n_A
AO			$2pr$	n^*_{AO}=[2r/(p+2r)] n_A
BB	B	n_B	q^2	n^*_{BB}=[q/(q+2r)] n_B
BO			$2qr$	n^*_{BO}=[2r/(q+2r)] n_B
AB	AB	n_{AB}	$2pq$	n_{AB}
OO	O	n_O	r^2	n_O

$$p'' = \frac{1}{2n}(2n^*_{AA} + n^*_{AO} + n_{AB}) \quad = \frac{p'+r'}{p+2r'}\frac{n_A}{n} + \frac{n_{AB}}{2n}$$

$$q'' = = \frac{q'+r'}{q+2r'}\frac{n_A}{n} + \frac{n_{AB}}{2n} \qquad r'' = \frac{r'}{p+2r'}\frac{n_A}{n} + \frac{r'}{q'+2r'}\frac{n_B}{n} + \frac{2n_O}{2n}$$

$$p' = 1 - \sqrt{(n_O + n_B)/n} \qquad q' = 1 - \sqrt{(n_O + n_A)/n} \qquad r' = \sqrt{n_O/n}$$

Example 2: Gametic frequencies: P_{AB}

$$P'_{Ab} = \tilde{P}_A - P'_{AB} \qquad P'_{aB} = \tilde{P}_B - P'_{AB}$$

$$P'_{ab} = 1 - \tilde{P}_A - \tilde{P}_B + P'_{AB}$$

$$P^{AB'}_{ab} = \frac{2p'_{AB}p'_{ab}}{2p'_{AB}p'_{ab} + 2p'_{Ab}p'_{aB}}\tilde{P}_{AaBb}$$

$$P^{Ab'}_{aB} = \frac{2p'_{Ab}p'_{aB}}{2p'_{AB}p'_{ab} + 2p'_{Ab}p'_{aB}}\tilde{P}_{AaBb}$$

$$P''_{AB} = \tilde{P}^{AB}_{AB} + \frac{1}{2}(\tilde{P}^{AB}_{Ab} + \tilde{P}^{AB}_{aB} + \frac{2p'_{AB}p'_{ab}}{2p'_{AB}p'_{ab} + 2p'_{Ab}p'_{aB}}\tilde{P}_{AaBb})$$

Within-population inbreeding coefficient

$$f = \frac{1}{n}\sum_u x_u$$

Where $x_u = \dfrac{fn_{uu}}{f + P_u(1-f)}$

Method of moments

$$\hat{f} = 1 - \frac{(n-1)n_{Aa}/n}{2n\tilde{P}_A\tilde{P}_a - n_{Aa}/2n}$$

Bayesian estimation

$$\Pr(B \mid A) = \frac{P(A,B)}{\Pr(A)}$$

Summary

Homework: Exercise 2.1 (page 88)