ALLELE FREQUENCIES

Binomial Distribution

The binomial probability of x successes in n trials is

$$\text{Pr}(x|p) \;=\; \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

The same quantity, written as $L(p|x)$, is the *likelihood of the parameter*, p, when the value x has been observed. The terms that do not involve p are not needed, so

$$L(p|x) \;\propto\; p^x(1-p)^{(n-x)}$$

Normal Approximation

Provided np is not too small the binomial distribution can be approximated by the normal distribution with the same mean and variance. In particular:

$$\tilde{p} \;\sim\; N\left(p,\ \frac{p(1-p)}{n}\right)$$

The standard normal variable z is

$$z \;=\; \frac{\tilde{p}-p}{\sqrt{p(1-p)/n}}$$

and 95% of z-values lie in

$$p \;\pm\; 1.96\sqrt{p(1-p)/n}$$

A 95% confidence interval for the binomial parameter p is

$$\tilde{p} \;\pm\; 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$$

Multinomial Distribution

If $\{P_i\}$ are the probabilities for a series of categories, the probability for counts $\{n_i\}$ is

$$\mathsf{Pr}(\{n_i\}) \;=\; \frac{n!}{\Pi_i n_i!}\Pi_i (P_i)^{n_i}$$

The sample proportions $\tilde{P}_i = n_i/n$ have these moments:

$$\begin{aligned}
\mathcal{E}(\tilde{P}_i) &= P_i \\
\mathsf{Var}(\tilde{P}_i) &= \frac{1}{n}P_i(1 - P_i) \\
\mathsf{Cov}(\tilde{P}_i, \tilde{P}_j) &= -\frac{1}{n}P_i P_j, \;\; i \neq j
\end{aligned}$$

Genotype and Allele Counts

The set of genotype counts $\{n_{AA}, n_{AB}, n_{BB}\}$ are multinomially distributed. The individual genotype counts $(n_{AA}, n - n_{AA})$ are binomially distributed.

The allele counts $n_A = 2n_{AA} + n_{AB}$ and $n_B = 2n_{BB} + n_{AB}$ are not binomially distributed unless there is Hardy-Weinberg equilibrium:

$$\text{Var}(\tilde{p}_A) = \frac{1}{2n}[p_A(1 - p_A) + (P_{AA} - p_A^2)]$$

Within-population Inbreeding Coefficient

Reparameterize genotype frequencies with the within-population *inbreeding coefficient* f:

$$\begin{aligned} P_{AA} &= p_A^2 + f p_A p_B \\ P_{AB} &= 2p_A p_B - 2f p_A p_B \\ P_{BB} &= p_B^2 + f p_A p_B \end{aligned}$$

$$\max\left(-\frac{p_A}{1-p_A}, -\frac{1-p_A}{p_A}\right) \leq \ f \ \leq 1$$

Maximum Likelihood Estimation of f

If $\tilde{p}_l$ is the sample frequency for the reference allele at SNP l, the MLEs for p_l and f are:

$$\begin{aligned} \hat{p}_l &= \tilde{p}_l \\ \hat{f} &= 1 - \frac{\tilde{H}_l}{2\tilde{p}_l(1 - \tilde{p}_l} \end{aligned}$$

where $\tilde{H}_l$ is the sample proportion of heterozygotes for SNP l.

This MLE has mean and variance

$$\begin{aligned} \mathcal{E}(\hat{f}) &\approx f \\ \mathrm{Var}(\hat{f}) &\approx \frac{1}{n}, \text{ if } f = 0 \end{aligned}$$

The bias of $\hat{f}$ is reduced by using large numbers of SNPs, as shown in Section 4.

EM Algorithm for Gamete Frequencies

There are nine distinguishable two-locus counts:

	BB	Bb	bb	Total
AA	n_{AABB}	n_{AABb}	n_{AAbb}	n_{AA}
Aa	n_{AaBB}	n_{AaBb}	n_{Aabb}	n_{Aa}
aa	n_{aaBB}	n_{aaBb}	n_{aabb}	n_{aa}
Total	n_{BB}	n_{Bb}	n_{bb}	n

and there is one unknown gamete count $x = n_{AB}$ for AB:

	B	b	Total
A	$n_{AB} = x$	$n_{Ab} = n_A - x$	$n_A = 2n_{AA} + n_{Aa}$
a	$n_{aB} = n_B - x$	$n_{ab} = x + n_b - n_A$	$n_a = 2n_{aa} + n_{Aa}$
Total	$n_B = 2n_{BB} + n_{Bb}$	$n_b = 2n_{bb} + n_{Bb}$	$2n$

The EM equation for the MLE of x is

$$x' \;=\; 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{2n_{AB}n_{ab}}{2n_{AB}n_{ab} + 2n_{Ab}n_{aB}} n_{AaBb}$$

Breakout Group Tasks

For the following allele dosage data: estimate the allele frequency and inbreeding coefficient for each of the 5 SNPs.

	Individual									
SNP	1	2	3	4	5	6	7	8	9	10
rs10492936	0	0	0	0	0	0	0	0	0	0
rs10489589	2	0	0	0	1	1	0	0	0	0
rs10489588	2	0	0	1	0	0	0	1	0	0
rs4472706	2	2	2	2	1	1	2	1	2	1
rs4587514	2	2	2	2	2	1	1	2	1	1

1. Estimate the allele frequency and inbreeding coefficient for each of the 5 SNPs.

2. Estimate f with all the SNPs, using $\widehat{f} = 1 - \sum_l \tilde{H}_l / \sum_l 2\tilde{p}_l(1 - \tilde{p}_l)$.

3. Estimate the gamete frequency for the reference alleles for SNPs rs10489589 and rs10489588.

4. Estimate the gamete frequency for the reference alleles for SNPs rs10489588 and rs4472706.