Maximum Likelihood Estimation: Binomial

For a sample of n independently-sampled alleles, n_A of type A and $n_a = n - n_A$ of type a, the likelihood of p_A is

$$L(p_A) \;=\; C(p_A)^{n_A}(1-p_A)^{n-n_A}$$

and this is maximized when $p_A = n_A/n$. The maximum likelihood estimate (MLE) of p_A is its sample value:

$$\widehat{p}_A \;=\; \tilde{p}_A$$

Aside: MLE Details

The likelihood function $L(p_A)$ is maximized by setting to zero its derivative with respect to p_A:

$$\frac{\partial L(p_A|n_A)}{\partial p_A} = 0 \quad \text{or when} \quad \frac{\partial \ln L(p_A|n_a)}{\partial p_A} = 0$$

Now

$$\ln L(p_A|n_A) \;=\; \ln C + n_A \ln(p_A) + (n - n_A)\ln(1 - p_A)$$

so

$$\frac{\partial \ln L(p_A|n_A)}{\partial p_A} \;=\; \frac{n_A}{p_A} - \frac{n - n_A}{1 - p_A}$$

and this is zero when $p_A = n_A/n$. The MLE of p_A is its sample value: $\widehat{p}_A = \tilde{p}_A$.

Maximum Likelihood Estimation: Multinomial

If $\{n_i\}$ are multinomial with parameters n and $\{P_i\}$, then the MLE's of P_i are n_i/n. This will always hold for genotype proportions, but not always for allele proportions.

For two alleles, the MLE's for genotype proportions are:

$$\begin{aligned} \hat{P}_{AA} &= n_{AA}/n \\ \hat{P}_{Aa} &= n_{Aa}/n \\ \hat{P}_{aa} &= n_{aa}/n \end{aligned}$$

Does this lead to estimates of allele proportions and the within-population inbreeding coefficient?

Maximum Likelihood Estimation

Because

$$\begin{aligned} P_{AA} &= p_A^2 + f p_A(1-p_A) \\ P_{Aa} &= 2p_A(1-p_A) - 2f p_A(1-p_A) \\ P_{aa} &= (1-p_A)^2 + f p_A(1-p_A) \end{aligned}$$

the likelihood function for p_A, f is

$$\begin{aligned} L(p_A, f) &= C[p_A^2 + p_A(1-p_A)f]^{n_{AA}} \\ &\quad \times [2p_A(1-p_A)f]^{n_{Aa}} [(1-p_A)^2 + p_A(1-p_A)f]^{n_{aa}} \end{aligned}$$

and it is difficult to find, algebraically, the values of p_A and f that maximize this function or its logarithm.

There is an alternative way of finding maximum likelihood estimates in this case: equating the observed and expected values of the genotype frequencies.

Bailey's Method

Because the number of parameters (2) equals the number of degrees of freedom in this case, we can just equate observed and expected genotype proportions based on the estimates of p_A and f:

$$\begin{aligned} n_{AA}/n &= \hat{p}_A^2 + \hat{f}\hat{p}_A(1-\hat{p}_A) \\ n_{Aa}/n &= 2\hat{p}_A(1-\hat{p}_A) - 2\hat{f}\hat{p}_A(1-\hat{p}_A) \\ n_{aa}/n &= (1-\hat{p}_A)^2 + \hat{f}\hat{p}_A(1-\hat{p}_A) \end{aligned}$$

Solving these equations (e.g. by adding the first equation to half the second equation to give solution for $\hat{p}_A$ and then substituting that into one equation):

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n} = \tilde{p}_A$$

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

Aside: Three-allele Case

With three alleles, there are six genotypes and 5 df. To use Bailey's method, would need five parameters: 2 allele frequencies and 3 inbreeding coefficients. For example

$$\begin{aligned} P_{11} &= p_1^2 + f_{12}p_1p_2 + f_{13}p_1p_3 \\ P_{12} &= 2p_1p_2 - 2f_{12}p_1p_2 \\ P_{22} &= p_2^2 + f_{12}p_1p_2 + f_{23}p_2p_3 \\ P_{13} &= 2p_1p_3 - 2f_{13}p_1p_3 \\ P_{23} &= 2p_2p_3 - 2f_{23}p_2p_3 \\ P_{33} &= p_3^2 + f_{13}p_1p_3 + f_{23}p_2p_3 \end{aligned}$$

We would generally prefer to have only one inbreeding coefficient f. It is a difficult numerical problem to find the MLE for f.

Method of Moments

An alternative to maximum likelihood estimation is the method of moments (MoM) where observed values of statistics are set equal to their expected values regardless of degrees of freedom. In general, this does not lead to unique estimates or to estimates with variances as small as those for maximum likelihood.

Bailey's method is for the special case where the MLEs are also MoM estimates.

Aside: MoM for Multiple Alleles

For the inbreeding coefficient at loci with m alleles A_u, two possible MoM estimates are (for large sample sizes)

$$\hat{f}_W = \frac{\sum_{u=1}^{m}(\tilde{P}_{uu} - \tilde{p}_u^2)}{\sum_{u=1}^{m}\tilde{p}_u(1-\tilde{p}_u)}$$

$$\hat{f}_H = \frac{1}{m-1}\sum_{u=1}^{m}\left(\frac{\tilde{P}_{uu} - \tilde{p}_u^2}{\tilde{p}_u}\right)$$

These both have low bias. Their variances depend on the value of f.

For loci with two alleles, $m = 2$, the two moment estimates are equal to each other and to the maximum likelihood estimate:

$$\hat{f}_W = \hat{f}_H = 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a}$$

MLE for Recessive Alleles

Suppose allele a is recessive to allele A, and a sample of n individuals has n_{aa} recessive homozygotes. The genotypes of the other $(n-n_{aa})$ individuals can be AA or Aa. If there is Hardy-Weinberg equilibrium, the likelihood for the two phenotypes is

$$\begin{aligned} L(p_a) &= (p_a^2)^{n_{aa}}(1-p_a^2)^{n-n_{aa}} \\ \ln[L(p_a)] &= 2n_{aa}\ln(p_a) + (n-n_{aa})\ln(1-p_a^2) \end{aligned}$$

Differentiating wrt p_a:

$$\frac{\partial \ln L(p_a)}{\partial p_a} = \frac{2n_{aa}}{p_a} - \frac{2p_a(n-n_{aa})}{1-p_a^2}$$

Setting this to zero leads to an equation that can be solved explicitly: $p_a = \sqrt{n_{aa}/n}$.

Aside: EM Algorithm for Recessive Alleles

An alternative way of finding maximum likelihood estimates when there are "missing data" involves *Estimation* of the missing data and then *Maximization* of the likelihood.

For a locus with allele A dominant to a the missing information is the counts of the AA and Aa genotypes. Only the joint count $(n - n_{aa})$ of $AA + Aa$ is observed.

Estimate the missing genotype counts (assuming independence of alleles) as proportions of the total count of dominant phenotypes:

$$n_{AA} = \frac{(1-p_a)^2}{1-p_a^2}(n-n_{aa}) = \frac{(1-p_a)(n-n_{aa})}{(1+p_a)}$$

$$n_{Aa} = \frac{2p_a(1-p_a)}{1-p_a^2}(n-n_{aa}) = \frac{2p_a(n-n_{aa})}{(1+p_a)}$$

Aside: EM Algorithm for Recessive Alleles

Maximize the likelihood (using Bailey's method):

$$\begin{aligned}
\widehat{p}_a &= \frac{n_{Aa} + 2n_{aa}}{2n} \\
&= \frac{1}{2n}\left(\frac{2p_a(n - n_{aa})}{(1 + p_a)} + 2n_{aa}\right) \\
&= \frac{2(np_a + n_{aa})}{2n(1 + p_a)}
\end{aligned}$$

An initial estimate p_a is put into the right hand side to give an updated estimated $\widehat{p}_a$ on the left hand side. This is then put back into the right hand side to give an iterative equation for p_a.

This procedure also has explicit solution $\widehat{p}_B = \sqrt{n_{aa}/n}$.

EM Algorithm for Two Loci

An interesting application of the EM algorithm is the estimation of two-locus gamete frequencies from unphased genotype data. For locus **A** with alleles A, a and locus **B** with alleles B, b, the ten two-locus frequencies are:

Genotype	Actual	Expected	Genotype	Actual	Expected
AB/AB	P_{AB}^{AB}	p_{AB}^2	AB/Ab	P_{Ab}^{AB}	$2p_{AB}p_{Ab}$
AB/aB	P_{aB}^{AB}	$2p_{AB}p_{aB}$	AB/ab	P_{ab}^{AB}	$2p_{AB}p_{ab}$
Ab/Ab	P_{Ab}^{Ab}	p_{Ab}^2	Ab/aB	P_{aB}^{Ab}	$2p_{Ab}p_{aB}$
Ab/ab	P_{ab}^{Ab}	$2p_{Ab}p_{ab}$	aB/aB	P_{aB}^{aB}	p_{aB}^2
aB/ab	P_{ab}^{aB}	$2p_{aB}p_{ab}$	ab/ab	P_{ab}^{ab}	p_{ab}^2

EM Algorithm for Two Loci

Gamete frequencies are marginal sums:

$$\begin{aligned} p_{AB} &= P_{AB}^{AB} + \frac{1}{2}(P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}) \\ p_{Ab} &= P_{Ab}^{Ab} + \frac{1}{2}(P_{AB}^{Ab} + P_{ab}^{Ab} + P_{aB}^{Ab}) \\ p_{aB} &= P_{aB}^{aB} + \frac{1}{2}(P_{AB}^{aB} + P_{ab}^{aB} + P_{Ab}^{aB}) \\ p_{ab} &= P_{ab}^{ab} + \frac{1}{2}(P_{Ab}^{ab} + P_{aB}^{ab} + P_{AB}^{ab}) \end{aligned}$$

Arrange the gamete frequencies as a two-way table to show that only one of them is unknown when the allele frequencies are known:

p_{AB}	p_{Ab}	p_A
p_{aB}	p_{ab}	p_a
p_B	p_b	1

EM Algorithm for Two Loci

The two double heterozygote counts n_{ab}^{AB}, n_{aB}^{Ab} are "missing data."

Assume initial value of p_{AB} and *Estimate* the missing counts as proportions of the total count n_{AaBb} of double heterozygotes:

$$\begin{aligned} n_{ab}^{AB} &= \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\ n_{aB}^{Ab} &= \frac{2p_{Ab}p_{aB}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \end{aligned}$$

and then *Maximize* the likelihood by setting

$$\begin{aligned} p_{AB} &= \frac{1}{2n}\left(2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB} + n_{ab}^{AB}\right) \\ &\text{or} \\ n_{AB} &= 2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB} + n_{ab}^{AB} \end{aligned}$$

Example

As an example, consider the data for two SNPs:

	BB	Bb	bb	Total
AA	$n_{AABB} = 0$	$n_{AABb} = 0$	$n_{AAbb} = 2$	$n_{AA} = 2$
Aa	$n_{AaBB} = 1$	$n_{AaBb} = 3$	$n_{Aabb} = 4$	$n_{Aa} = 8$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 1$	$n_{aabb} = 4$	$n_{aa} = 5$
Total	$n_{BB} = 1$	$n_{Bb} = 4$	$n_{bb} = 10$	$n = 15$

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

	B	b	Total
A	$n_{AB} = x$	$n_{Ab} = 12 - x$	$n_A = 12$
a	$n_{aB} = 6 - x$	$n_{ab} = x + 12$	$n_a = 18$
Total	$n_B = 6$	$n_b = 24$	$2n = 30$

$$0 \leq x \leq 6$$

Example

EM iterative equation:

$$
\begin{aligned}
x' &= 2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AB/ab} \\
&= 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\
&= 0 + 0 + 1 + 3 \times \frac{2x(x+12)}{2x(x+12) + 2(12-x)(6-x)} \\
&= 1 + \frac{3x(x+12)}{x(x+12) + (12-x)(6-x)}
\end{aligned}
$$

Example

A good starting value would assume independence of A and B alleles: $x = 2n * p_A * p_B = (30 \times 12/30 \times 6/30) = 2.4$. Successive iterates are:

Iterate	x	$x/2n$
1	2.4000	0.0800
2	2.5000	0.0833
3	2.5647	0.0855
4	2.6063	0.0869
5	2.6327	0.0878
6	2.6494	0.0883
7	2.6600	0.0887
8	2.6667	0.0889
9	2.6709	0.0890
10	2.6736	0.0891
11	2.6752	0.0892
12	2.6763	0.0892
13	2.6769	0.0892
14	2.6773	0.0892
15	2.6776	0.0893
16	2.6778	0.0893
...	...	...