

## II. Population Genetics

# Maximum Likelihood Estimation for Allele Frequencies

---

### Lecture 15

Instructor: Su-In Lee

## Last Lectures: Introduction to Coalescent Model

---

- Computationally efficient simulations
  - Alternative to a forward approach
- Predictions about sequence variation
  - Number of polymorphisms
  - Frequency of polymorphisms

## Coalescent Models: Key Ideas

---

- Proceed backwards in time
- Genealogies shaped by
  - Population size
  - Population structure
  - Recombination rates
- Given a particular genealogy ...
  - Mutation rate predicts variation

## What Next?

---

- Estimating allele and haplotype frequencies from genotype data
  - Maximum likelihood approach
  - Application of an E-M algorithm
- Challenges
  - Using information from related individuals
  - Allowing for non-codominant genotypes
  - Allowing for ambiguity in haplotype assignments

## Objective: Parameter Estimation

---

- Learn about **population characteristics**
  - E.g. allele frequencies, population size
- Using a specific **sample**
  - E.g. a set sequences, unrelated individuals, or even families

## Maximum Likelihood

---

- A general framework for estimating model parameters
- Find the set of parameter values that maximize the probability of the observed data
- Applicable to many different problems

## Example: Allele Frequencies

- Consider...
  - A sample of  $n$  chromosomes
  - $X$  of these are of type “a”
  - Parameter of interest is allele frequency...

$$L(p | n, X) = \binom{n}{X} p^X (1-p)^{n-X}$$

p	1-p	L
0.0	1.0	0.000
0.2	0.8	0.088
0.4	0.6	0.251
0.6	0.4	0.111
0.8	0.2	0.006
1.0	0.0	0.000

## In this case

- The likelihood tells us the data is most probable if  $p = 0.4$
- The likelihood curve allows us to evaluate alternatives...
  - Is  $p = 0.8$  a possibility?
  - Is  $p = 0.2$  a possibility?

## Another Example: Estimating $4N\mu$

- Consider  $S$  polymorphisms in sample of  $n$  sequences...

$$L(\theta | n, S) = P_n(S | \theta),$$

- where  $P_n$  is calculated using the  $Q_n$  and  $P_2$  functions defined previously

## Maximum Likelihood Estimation

- Two steps
  - Write down likelihood function  $L(\theta | x) \propto f(x | \theta)$
  - Find values of  $\hat{\theta}$  that maximizes  $L(\theta | x)$
- In principle, applicable to any problem where a likelihood function exists

## Analytical Solutions

---

- Write out log-likelihood

$$l(\theta | data) = \log L(\theta | data)$$

- Calculate derivative of likelihood

$$\frac{dl(\theta | data)}{d\theta}$$

- Find zeros for derivative function

## Allele Frequency Estimation

---

- When individual chromosomes are observed this does not seem tricky...
- What about with genotypes?
- What about with parent-offspring pairs?

## Coming up ...

- We will walk through allele frequency estimation in three distinct settings:
  - Samples single chromosomes ...
  - Samples of unrelated Individuals ...
  - Samples of parents and offspring ...

## 1. Single Alleles Observed

- Consider
  - A sample of  $n$  chromosomes
  - $X$  of these are type “a”
  - Parameter of interest is allele frequency ...

$$L(p | n, X) = \binom{n}{X} p^X (1-p)^{n-X}$$

- The following two likelihoods are just as good:

$$L(p | n, X) = \binom{n}{X} p^X (1-p)^{n-X}$$

$$L(p | x_1, \dots, x_n) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

- For ML estimation, constant factors in likelihood don't matter

## Analytic Solution

- The log-likelihood

$$\log L(\theta | n, X) = \log \binom{n}{X} + X \log p + (n - X) \log(1 - p)$$

- The derivative

$$\frac{d \log L(p | X)}{dp} = \frac{X}{p} - \frac{n - X}{1 - p}$$

- Find zero ...

## Samples of Individual Chromosomes

- The natural estimator (where we count the proportion of sequences of a particular type) and the MLE give identical solutions
- Maximum likelihood provides a justification for using the “natural” estimator

## 2. Genotypes Observed

Genotypes				
Genotype	$A_1A_1$	$A_1A_2$	$A_2A_2$	Total
Observed	$n_{11}$	$n_{12}$	$n_{22}$	$n = n_{11} + n_{12} + n_{22}$
Frequency	$p_{11}$	$p_{12}$	$p_{22}$	1.0

  

Alleles			
Genotype	$A_1$	$A_2$	Total
Observed	$n_1 = 2n_{11} + n_{12}$	$n_2 = 2n_{22} + n_{12}$	$2n = n_1 + n_2$
Frequency	$p_1 = n_1 / 2n$	$p_2 = n_2 / 2n$	1.0

- Log-likelihood and its derivative

$$l = \log L = (2n_{11} + n_{12}) \log p_1 + (2n_{22} + n_{12}) \log(1 - p_1) + C$$

$$\frac{dl}{dp_1} = \frac{2n_{11} + n_{12}}{p_1} - \frac{2n_{22} + n_{12}}{1 - p_1}$$

- Giving the MLE as ...

$$\hat{p}_1 = \frac{(2n_{11} + n_{12})}{2(n_{11} + n_{12} + n_{22})}$$

## Samples of Unrelated Individuals

- Again, natural estimator (where we count the proportion of alleles of a particular type) and the MLE give identical solutions
- Maximum likelihood provides a justification for using the “natural” estimator

### 3. Parent-Offspring Pairs

Parent	Child			
	$A_1A_1$	$A_1A_2$	$A_2A_2$	
$A_1A_1$	$a_1$	$a_2$	0	$a_1+a_2$
$A_1A_2$	$a_3$	$a_4$	$a_5$	$a_3+a_4+a_5$
$A_2A_2$	0	$a_6$	$a_7$	$a_6+a_7$
	$a_1+a_3$	$a_2+a_4+a_6$	$a_5+a_7$	N pairs

### Probability of Each Observation

Parent	Child			
	$A_1A_1$	$A_1A_2$	$A_2A_2$	
$A_1A_1$	$p_1^3$	$p_1^2p_2$	0	$p_1^2$
$A_1A_2$	$p_1^2p_2$	$p_1p_2$	$p_1p_2^2$	$2p_1p_2$
$A_2A_2$	0	$p_1p_2^2$	$p_2^3$	$p_2^2$
	$p_1^2$	$2p_1p_2$	$p_2^2$	1.0

- Giving the MLE as

$$\begin{aligned}\log L &= a_1 \log p_1^3 + (a_2 + a_3) \log(p_1^2 p_2) + a_4 \log(p_1 p_2) \\ &\quad + (a_5 + a_6) \log(p_1 p_2^2) + a_7 \log(p_2^3) + \text{constant} \\ &= B \log p_1 + C \log(1 - p_1)\end{aligned}$$

$$B = 3a_1 + 2(a_2 + a_3) + a_4 + (a_5 + a_6)$$

$$C = (a_2 + a_3) + a_4 + 2(a_5 + a_6) + 3a_7$$

$$\hat{p}_1 = \frac{B}{B + C}$$

## Samples of Parent Offspring-Pairs

- The natural estimator (where we count the proportion of alleles of a particular type) and the MLE no longer give identical solutions
- In this case, we expect the MLE to be more accurate

## Summary

---

- Examples of Maximum Likelihood Estimation in Genetics
- Allele Frequency Estimation
  - Allele counts
  - Genotype counts
  - Pairs of individuals

## Acknowledgement

---

- This lecture note is based on material by Profs Goncalo Abecasis (Univ of Michigan).