Maximum Likelihood (Part I)

Biostatistics 666
Properties of genes in populations

- Allele frequencies
- Genotype frequencies
  - Hardy-Weinberg Equilibrium
- Haplotype frequencies
  - Linkage Equilibrium
- Penetrance
- Degree of variation in a sequence ...
Recommended Reading...


- Used coalescent methods to predict correlation between neighboring polymorphisms in humans...
Parameter Estimation

- Population parameters are unknown
  - E.g. allele frequencies, population size

- Use a **sample** from a **population**
  - E.g. unrelated individuals, families

- Make inferences on population properties
Properties of Estimators

- Estimators can have several desirable properties, such as:
  - Unbiasedness
  - Consistency
  - Efficiency
  - Sufficiency

- Choosing among alternative estimators requires evaluating these
Unbiasedness

- An estimator is unbiased if

\[ E(\hat{\theta}) = \theta \]

\[ \text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta \]

- Multiple unbiased estimators may exist
- Other properties may be desirable
An estimator is consistent if

$$P\left( |\hat{\theta} - \theta | > \varepsilon \right) \to 0 \text{ as } n \to \infty$$

for any $\varepsilon$

Estimate converges to true value in probability with increasing sample size
Mean Squared Error

- MSE is defined as

\[ MSE(\hat{\theta}) = E\left(\left\{ \left(\hat{\theta} - \bar{\theta}\right) + (\bar{\theta} - \theta) \right\}^2 \right) \]

\[ = \text{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2 \]

- If \( MSE \to 0 \) as \( n \to \infty \) then the estimator must be consistent
  - The reverse is not true
Efficiency

- The relative efficiency of two estimator is the reciprocal or the ratio of their variances

\[
\text{if } \frac{\text{var}(\hat{\theta}_2)}{\text{var}(\hat{\theta}_1)} > 1 \text{ then } \hat{\theta}_1 \text{ is more efficient}
\]

- Comparison only meaningful for estimators with equal biases
Sufficiency

- Consider...
  - Observations $X_1, X_2, \ldots, X_n$
  - Statistic $T(X_1, X_2, \ldots, X_n)$

- $T$ is a sufficient statistic if it includes all information about $\theta$ in the sample
  - Distribution of $X_i$ conditional on $T$ is independent of $\theta$
  - Posterior distribution of $\theta$ conditional on $T$ is independent of $X_i$
Minimal Sufficient Statistic

- There can be many alternative sufficient statistics.

- A statistic is a minimal sufficient statistic if it can be expressed as a function of every other sufficient statistic.
Maximum Likelihood Estimation

- Widely used in genetics
- Two basic steps...
  
  a) Write down likelihood function

  \[ L(\theta; x) = f(x; \theta) \]

  b) Find value of \( \hat{\theta} \) that maximizes \( L(\theta; x) \)

- In principle, might be applicable to any problem where a likelihood function exists
MLEs

- The parameter values that maximize likelihood
  - $\theta$ for which observations have maximum probability

- Finding MLEs is an optimization problem
Typical Properties of MLEs

- **Bias**
  - Can be biased or unbiased

- **Consistency**
  - Subject to regularity conditions, MLEs are consistent

- **Efficiency**
  - Typically, MLEs are asymptotically efficient estimators

- **Sufficiency**
  - Often, but not always

- Cox and Hinkley, 1974
Example: Allele Frequencies

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

\[
L(p; x, n) = \binom{n}{X} p^x (1 - p)^{n-x}
\]
Likelihood Plot

For $n = 10$ and $X = 4$
Analytic Solution

- Sometimes possible, often uses log-likelihood

\[ \ln L(\theta; x, n) = \ln \binom{n}{X} + X \ln p + (n - X) \ln(1 - p) \]

- Calculate derivative

\[ \frac{d \ln L(p; x)}{dp} = \frac{X}{p} - \frac{n - X}{1 - p} \]

- Find zero ...
Genotype data is more common...

<table>
<thead>
<tr>
<th>Genotype</th>
<th>$A_1A_1$</th>
<th>$A_1A_2$</th>
<th>$A_2A_2$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{22}$</td>
<td>$n=n_{11}+n_{12}+n_{22}$</td>
</tr>
<tr>
<td>Frequency</td>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
<td>$p_{22}$</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotype</th>
<th>$A_1$</th>
<th>$A_2$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>$n_1=2n_{11}+n_{12}$</td>
<td>$n_2=2n_{22}+n_{12}$</td>
<td>$2n=n_1+n_2$</td>
</tr>
<tr>
<td>Frequency</td>
<td>$p_1=n_1/2n$</td>
<td>$p_2=n_2/2n$</td>
<td>1.0</td>
</tr>
</tbody>
</table>
MLE using genotype data...

Consider a sample such as ...

<table>
<thead>
<tr>
<th>Genotype</th>
<th>$A_1A_1$</th>
<th>$A_1A_2$</th>
<th>$A_2A_2$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{22}$</td>
<td>$n = n_{11} + n_{12} + n_{22}$</td>
</tr>
</tbody>
</table>

The likelihood as a function of allele frequencies is ...

$$L(p; n) = \frac{n!}{n_{11}! n_{12}! n_{22}!} \left( p^2 \right)^{n_{11}} \left( 2pq \right)^{n_{12}} \left( q^2 \right)^{n_{22}}$$
Which gives...

- Log-likelihood ...

\[ \ln L = (2n_{11} + n_{12}) \ln p_1 + (2n_{22} + n_{12}) \ln p_2 + C \]

- With derivative ...

\[ \frac{d \ln L}{dp_1} = \frac{(2n_{11} + n_{12})}{p_1} - \frac{(2n_{22} + n_{12})}{(1 - p_1)} \]

- Giving the MLE as ...

\[ \hat{p}_1 = \frac{(2n_{11} + n_{12})}{2(n_{11} + n_{12} + n_{22})} \]
Parent-Offspring Pairs

- Likelihoods can be written for more complex situations...

<table>
<thead>
<tr>
<th>Parent</th>
<th>Child</th>
<th>Parent Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A_1A_1$</td>
<td>$a_1$</td>
</tr>
<tr>
<td>$A_1A_1$</td>
<td>$a_2$</td>
<td>$A_1A_2$</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$a_3$</td>
<td>$a_4$</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$0$</td>
<td>$a_6$</td>
</tr>
</tbody>
</table>
Parent Offspring Pairs (II)

<table>
<thead>
<tr>
<th>Parent</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_{1}A_{1}</td>
<td>A_{1}A_{1} p_{1}^{3} A_{1}A_{2} p_{1}p_{2} A_{2}A_{2} 0 Parent Total p_{1}^{2}</td>
</tr>
<tr>
<td>A_{1}A_{2}</td>
<td>A_{1}A_{2} p_{1}^{2} p_{2} A_{1}A_{2} p_{1}p_{2} A_{2}A_{2} p_{1}p_{2}^{2} 2p_{1}p_{2}</td>
</tr>
<tr>
<td>A_{2}A_{2}</td>
<td>A_{2}A_{2} 0 A_{2}A_{2} p_{1}p_{2}^{2} A_{2}A_{2} p_{2}^{3} p_{2}^{2}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Offspring Total</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_{1}^{2}</td>
<td>2p_{1}p_{2}</td>
</tr>
<tr>
<td>2p_{1}p_{2}</td>
<td>p_{2}^{2}</td>
</tr>
<tr>
<td>p_{2}^{2}</td>
<td>1</td>
</tr>
</tbody>
</table>
Which gives...

\[
\ln L = a_1 \ln p_1^3 + (a_2 + a_3) \ln(p_1^2 p_2) + a_4 \ln(p_1 p_2) \\
+ (a_5 + a_6) \ln(p_1 p_2^2) + a_7 \ln p_2^3 + C
\]

\[
= B \ln p_1 + C \ln p_2
\]

\[
p_2 = 1 - p_1
\]

\[
B = 3a_1 + 2(a_2 + a_3) + a_4 + (a_5 + a_6)
\]

\[
C = (a_2 + a_3) + a_4 + 2(a_5 + a_6) + 3a_7
\]

\[
\hat{p}_1 = \frac{B}{B + C}
\]
Estimating $4N\mu$

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

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

- Where $P_n$ and $Q_n$ and $P_2$ were defined previously
With $S = 10$, $n = 5$
More examples...

- Estimating ABO locus allele frequencies using phenotypes

- Estimating haplotype frequencies using genotype data

- Estimating $\theta = 4N\mu$ using polymorphism data
An introduction to Maximum Likelihood and the properties of estimators is available in chapters 2 and 3 of...