

# Model-based Inference

In statistics, there are two frameworks for <sup>model-based</sup> inference:

Maximum Likelihood

Bayesian Analysis

"Frequentists"

"Bayesians"

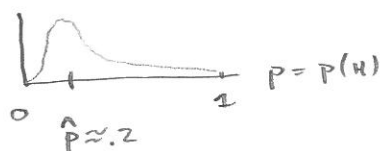
$p(H) = p$  unknown

get a POSTERIOR DIST.

H T H T T T H ...  $\rightarrow \infty$

on your parameter of interest

$\hat{p}$  = proportion of H  $\rightarrow p$   
in  $n$  trials



We begin with the Likelihood framework.

Warning: Need to review optimization.

Easy example: Show that the Maximum Likelihood Estimator (MLE)  $\hat{p}$

for  $n$  coin flips is  $\hat{p} = \frac{\# \text{ of H}}{n} \equiv \text{proportion of H in } n \text{ tosses}$

Setup: 1)  $p = p(H)$  is unknown and will be estimated

2) there are  $n$  coin flips and  $n_H$  are H,  $n_T$  are T,  $n_H + n_T = n$ .

The data are  $n_H, n_T \equiv \text{data}$ ,

With  $\underline{p}$  unknown, define the likelihood function

(Assumes i.i.d.)

$$L(p) = \text{Prob}(\text{data} \mid p) = p^{n_H} (1-p)^{n_T}$$

then the MAXIMUM LIKELIHOOD ESTIMATOR  $\hat{p} = \hat{p}_{MLE}$  is the value of  $p$  that maximizes  $L(p)$ . [The value of  $p$  that makes the data you observed most probable.]

Eg.  $n = 100$     $n_H = 30$     $n_T = 70$

$$L(p) = \text{Prob}(\text{data} | p) = p^{30} (1-p)^{70}$$

Quick Calc I optimization:

$$\begin{aligned} L'(p) &= p^{30} [70(1-p)^{69}(-1)] + 30p^{29}(1-p)^{70} \\ &= p^{29}(1-p)^{69} [30(1-p) - 70p] \end{aligned}$$

Require  $L'(p) = 0 \rightarrow$

$$p^{29} (1-p)^{69} [30 - 100p] = 0 \quad \text{i.e. } p=0, 1, \frac{30}{100} = .3$$

Using your favorite test, it is clear  $p = .3$  is a global max.

$$\therefore \hat{p} = \hat{p}_{MLE} = .3 = \frac{n_H}{n} \quad \square$$

Redo: Since ML will require set derivatives to zero (and typically iid assumption) in practice, use the Log-likelihood function.

Since  $\log(x)$  is an increasing function, the log-likelihood and likelihood function are maximized at the same value of  $p$ .

$$\begin{aligned} \text{log-likelihood} \equiv \ln L(p) &= \ln(p^{30} (1-p)^{70}) &= \ln p^{30} + \ln(1-p)^{70} \end{aligned}$$

Differentiate:  $\frac{d}{dp} \ln L(p) = \frac{1}{p^{30}} - \frac{1}{(1-p)^{70}}$

Easy to differentiate...

has critical points at  $p=0, 1, .3$

Caution: MLEs may fail to exist or not be unique.

# Maximum Likelihood Trees in phylogenetics

Goal: Fix a site substitution model JC, K2P, GTR

and find the parameters  $T, (\vec{p}, Q, \{t_i\})$  that maximize the likelihood

Example: Model Jukes-Cantor on 1-edge tree



$$\vec{p} = (.25, .25, .25, .25) \text{ known}$$

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ & \text{etc.} & & \end{pmatrix} \text{ known}$$

Unknown parameter  $t = \text{branch length}$

Data:  $F_{ab}$  = pattern frequencies counts of pattern (under iid assumption)

$$= \begin{pmatrix} n_{AA} & n_{AG} & n_{AC} & n_{AT} \\ n_{GA} & & & \\ n_{CA} & & & \\ n_{TA} & & & n_{TT} \end{pmatrix} \quad \begin{matrix} 16 \text{ bits of} \\ \text{data} \\ n_{ij} \end{matrix}$$

Under JC, the expected pattern frequency array is  $\begin{pmatrix} \frac{1-a(t)}{4} & \frac{a(t)}{12} & \frac{a(t)}{12} & \frac{a(t)}{12} \\ & \ddots & & \end{pmatrix}$

where  $a(t)$  is a <sup>invertible</sup> function of the unknown parameter  $t$

$$a(t) = a = \frac{3}{4} (1 - e^{-\frac{4}{3}t})$$

Thus, the log-likelihood is

$$\ln L(a) = \ln L(a | \text{data}) = \ln (\text{Prob}(\text{data} | a))$$

$$= \ln \begin{bmatrix} n_{AA} & n_{AG} & \dots & n_{TT} \\ p_{AA} & p_{AG} & & p_{TT} \end{bmatrix} = \ln \left( \prod_{\text{pattern } ij} P_{ij}^{n_{ij}} \right)$$

$$= \ln \begin{bmatrix} n_{AA} + n_{AG} + n_{AC} + n_{TT} & \sum_{i,j=1}^4 n_{ij} \\ p_{ii} & p_{ij} \end{bmatrix}$$

$$\ln L(a) = \ln \left[ \left( \frac{1-a}{3} \right)^{\sum n_{ii}} \left( \frac{a}{12} \right)^{\sum_{i \neq j} n_{ij}} \right] = \left( \sum n_{ii} \right) \ln \left( \frac{1-a}{3} \right) + \left( \sum n_{ij} \right) \ln \left( \frac{a}{12} \right)$$

Differentiating ...

$$\begin{aligned} \frac{d}{dt} \ln L(a(t)) &= \sum n_{ii} \left( \frac{3}{1-a} \right) \left( \frac{-1}{3} \right) a'(t) + \sum_{i \neq j} n_{ij} \frac{12}{a} \left( \frac{1}{12} \right) a'(t) \\ &= - \sum n_{ii} \frac{1}{1-a(t)} a'(t) + \sum_{i \neq j} n_{ij} \frac{1}{a} a'(t) \end{aligned}$$

Assume  $a'(t) = 0$  (i.e. there is change!) and require equal to zero

$$\frac{d}{dt} \ln L(a(t)) = 0 \Rightarrow \sum n_{ii} \frac{1}{1-a} = \sum n_{ij} \frac{1}{a}$$

$$\Rightarrow \left( \sum n_{ii} \right) a = \left( \sum n_{ij} \right) - \left( \sum n_{ij} \right) a \Rightarrow a = \frac{\sum n_{ij}}{\sum n_{ii} + \sum n_{ij}}$$

$$\text{i.e. } \hat{a}(t) = \frac{\sum n_{ij}}{n} = \text{proportion of sites that differ!}$$

$$\text{and } \hat{t} = \hat{t}_{MLE} = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} \hat{a} \right) = d_{JC}(a, b)$$

i.e. The JC distance is the Maximum Likelihood estimate for  $t$ !