

## Exercise 5: Statistical inference (II)

Siyue Yang

05/21/2022

### EM and Newton-Raphson implementation

The ABO-gene or ABO-locus is on chromosome 9. It has 3 alleles (antigens) ( $A, B, O$ ) and it determines 4 blood type ( $A, B, AB, O$ ).

genotype	phenotype
AA AO	A
BB BO	B
AB	AB
OO	O

A, B are dominant to O.  
O is recessive to A, B.  
A, B are co-dominant.

We have a large random sample obtained from Berlin (Bernstein 1925, Sham's book page 44):

- $n_A = 9123$  blood type A
- $n_B = 2987$  blood type B
- $n_{AB} = 1269$  blood type AB
- $n_O = 7725$  blood type O

For instance,  $n_A = 9123 = n_{AA} + n_{AO}$  : Among 9123 blood type A individuals, some have genotype AA and the others have genotype AO.

Our interest is to estimate the allele frequencies of alleles A, B, and O. i.e.  $p = \text{freq}(\text{allele } A)$ ,  $q = \text{freq}(\text{allele } B)$ ,  $1 - p - q = \text{freq}(\text{allele } O)$ .

1. Write out the log-likelihood  $L(p, q)$ .
2. Is there a closed-form solution of this log-likelihood function?
3. Formulate the problem as a missing data problem and use the Newton-Raphson algorithm to find the MLEs,  $\hat{p}$  and  $\hat{q}$ , that maximize the log-likelihood,  $\ln L(p, q)$ .
4. (Advanced) Use the EM algorithm to find the Maximum Likelihood Estimates (MLEs) of parameters,  $\hat{p}$  and  $\hat{q}$ .

Hint: Lei Sun's STA2080 Modern genetic statistics notes ([link](#)).