Biostatistics 666
Problem Set 3

Gene deletions and duplications.

Scientists studying the GSTM1 gene, located on chromosome 1, noted that because of gene deletions and duplications each chromosome could carry 0, 1 or 2 functional copies of the gene. In this way, a diploid individual could carry between 0 (corresponding to two deletion chromosomes) and 4 copies of the gene (corresponding to two duplication chromosomes).

Suppose an assay is available to estimate the total number of gene copies in an individual, between 0 and 4.

a) Given the three alleles (deletion, wild-type and duplication), what are the possible genotypes at the locus? Does each genotype correspond to a unique "phenotype" or assay result?

The six possible genotypes are: 0/0, 0/1, 0/2, 1/1, 1/2, 2/2. In this list, each digit represents the number of copies on one particular chromosome.

There are only 5 possible assay results (0, 1, 2, 3, 4), so not all genotypes correspond to unique "phenotype". In particular, genotypes 1/1 and 0/2 result in the same total copy number of 2.

b) Suppose you want to estimate allele frequencies for the deletion, wild-type and duplication alleles (p_0, p_1 and p_2). Specify an appropriate likelihood for studying these frequencies, using the total number of GSTM1 copies in each individual as input.

$$L = (p_0^2)^{n_0}(2p_0p_1)^{n_1}(p_1^2 + 2p_0p_2)^{n_2}(2p_1p_2)^{n_3}(p_2^2)^{n_4}$$

c) The E-M algorithm is often a convenient strategy for allele frequency estimation. Suppose an E-M algorithm where used to iteratively estimate allele frequencies at this locus. Describe how allele frequency estimates would be updated at each iteration, including appropriate formulae.

To estimate allele frequencies, we would have to distribute individuals with 2 copies between the two possible configurations $n_{0/2}$ and $n_{1/1}$, like this:

$$E(n_{11}^t) = n_2 P(genotype\ is\ 1/1\ |copy\ number\ is\ 2)$$

$$E(n_{11}^t) = n_2 \frac{P(genotype\ is\ 1/1)}{P(copy\ number\ is\ 2)}$$

$$E(n_{11}^t) = n_2 \frac{(p_1^t)^2}{(p_1^t)^2 + 2p_0^t p_2^t}$$

$$E(n_{02}^t) = n_2 P(genotype\ is\ 0/2\ |copy\ number\ is\ 2)$$

$$E(n_{02}^t) = n_2 \frac{P(genotype\ is\ 0/2)}{P(copy\ number\ is\ 2)}$$

$$E(n_{02}^t) = n_2 \frac{2p_0^t p_2^t}{(p_1^t)^2 + 2p_0^t p_2^t}$$

None of the other copy numbers are ambiguous (for example, copy number 0 always corresponds to genotype 0/0, copy number 1 corresponds to genotype 0/1, copy number 3 corresponds to genotype 1/2, and copy number 4 corresponds to genotype 2/2.

Once we have estimated the number of individuals with each genotype, we can update allele frequencies as:

$$p_0^{t+1} = \frac{2E(n_{00}) + E(n_{01}) + E(n_{02})}{2n} = \frac{2n_0 + n_1 + E(n_{02}^t)}{2n}$$
$$p_1^{t+1} = \frac{2E(n_{11}) + E(n_{01}) + E(n_{12})}{2n} = \frac{2E(n_{11}^t) + n_1 + n_3}{2n}$$
$$p_2^{t+1} = \frac{2E(n_{22}) + E(n_{12}) + E(n_{02})}{2n} = \frac{2n_4 + n_3 + E(n_{02}^t)}{2n}$$

d) How would you verify that the E-M algorithm converged to a maximum likelihood solution?

One possibility would be to calculate the second derivative (it should be negative if we have found a maximum in the likelihood). Further re-assurance would come from re-starting the estimation process from a different starting point and verifying we obtain the same solution (that would re-assure us we are not at a local maximum). If other parameter optimization techniques are available, we could check that they provide similar estimates for the MLE.

e) How would you estimate confidence intervals for your estimated allele frequencies?

One strategy would be to use the second derivative to calculate information and standard errors for our parameter estimates (this would assume that they are normally distributed). Then, using normal distribution quantiles, we could estimate a confidence interval.

Another option, would be to calculate the range of values of our parameters for which the log-likelihood ratio is relatively close to its maximum value. Given a desired confidence level, we could find the range of values that would not be rejected with our target confidence by a likelihood ratio test.

Other general purpose strategies (such as the bootstrap, which we did not discuss) are possible.