Biostatistics 666
Problem Set 3
Due October 11, 2017

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?

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 alleles in each individual as input.

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.

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

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