Statistical Genetics, Spring 2024

Homework exercise 2

Due date: 24/7 before class

1. **Phylogenetic reconstruction and hypothesis testing**
   This problem uses the following resources:

   - The program PHYLIP — this program is really easy to download and install, and programs are run simply by clicking the executable).
   - The 14-species primates+mammals mtDNA database available from PHYLIP. This file is already in PHYLIP format, no processing required. See documentation.

   (a) Estimate the phylogeny of the sample by running maximum parsimony (program dnapars) and by maximum likelihood (program dnaml), both with the default parameters. Consider the resulting trees in the file outfile (it is initially generated in the same directory as the executable). Comment on the differences.

   (b) Try generating a different phylogeny by playing with the parameter T (Transition/transversion ratio) in dnaml. Do you get a different result with T=1 or T=100?

   (c) Use the program dnamlk to generate a tree under the molecular clock assumption. Does the phylogeny make sense based on your biological knowledge? Perform a likelihood ratio test to determine appropriateness of molecular clock. For help on how to do this you may look at the dnamlk help page and read the paragraph starting:

   > "This program makes possible a (reasonably) legitimate statistical test of the molecular clock."

2. **Measures of LD on HapMap data**
   This problem uses the HapMap Yoruban haplotype data on Chromosome 22. Each row is a SNP, with its name, location on the chromosome, and the value of this SNP on all HapMap Yoruban haplotypes.

   (a) Pick 10000 values equally spread between 1 and 10^7. For each value x, pick at random two SNPs that are about x apart on the chromosome, and calculate their $|D'|$ and R^2 LD values. Plot a sample of 200 $|D'|$ and R^2 values as a function of distance (use log scale or other transformations as needed). Comment on the monotonicity of the graphs – which one appears more monotone?.

   (b) Model each relationship as a noisy curve using linear regression with appropriate transformations of the distance. Comment on the results, and on the appropriateness of the linear regression inference for this problem.

   (c) Model LD (either measure) as a function of both distance and location along the chromosome. Does the location have a significant effect on the LD? Interpret the results in terms of recombination rates.

3. **Two-stage designs: power and multiple comparisons**
   In this problem, we will convince ourselves that two-stage designs lose power, and examine the different ways of correcting for multiple comparisons. Assume we have $n = 500,000$ independent tests based on

$X_1, ..., X_n$ with $X_i \sim N(\mu_i, 100/m)$, where m is the number of individuals tested. For each observation, we want to test $H_0 : \mu_i = 0$ vs $H_1 : \mu_i = 1$. Assume we use simple Bonferroni corrections at level $\alpha = 0.05$ for multiple comparisons.

(a) For $m = 1000$, calculate the power of each test when correcting for $500,000$ comparison.

(b) Consider the alternative approach of using a random subset of size $m_0 = 500$, setting an initial p-value threshold of 0.001 and for the roughly 500 tests that are expected to pass this threshold, performing a follow-up analysis on the remaining $m_1 = 500$ subjects.

   i. Assuming we perform a second test on the follow up data only, correcting for (exactly!) 500 comparisons, prove that the overall probability of a false rejection is indeed bounded by α, even though more than 500 can pass the first threshold (in formal multiple comparisons speak, this means FWE is controlled at level α).
   Calculate the power of this approach.

   ii. Now assume that for the chosen tests, we now perform a test combining the data from the two stages, but correcting for $500,000$ comparisons. Prove that this also controls FWE at level α. Prove that this approach has lower power than the one-step approach in item (a).

   iii. (* Extra credit) Calculate explicitly the power of this last approach, and compare to the one from item (b)i.

(c) (*Lots of extra credit) Prove or disprove the following: the two stage method of item 2(b)(i) (set threshold p for first stage on half of samples, and correct for $p \times n$ in second stage on the second half) always has inferior power compared to the one-stage procedure of testing all n hypotheses on the complete sample (make explicit the conditions and assumptions you use).