Biostatistics 666
Sample Mid-Term Assessment

Instructions

Please read these carefully before proceeding with the exam.

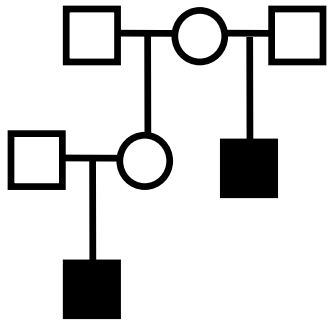
1. Please write **your name** in every page of your answer sheets.

2. You will be graded for **three of the five** problems in this exam. You can choose to answer any three.

3. Please present **all formulae and intermediate** calculations in your answers.

PROBLEM 1.

You are studying the relationship between genetic variation and blood iron levels in a sample of many small pedigrees. One of the pedigrees being studied is illustrated below:



a) In the pedigree illustrated above, what is the kinship coefficient between the two shaded individuals?

b) With appropriate genotype data, this sample of pedigrees could be used to study genetic linkage or association. What is the difference between a genetic *linkage* test and a genetic *association* test?

c) Suppose you have just received genetic marker data for all individuals in your sample. Outline three important quality checks you'd carry out before proceeding to linkage and/or association analyses.

d) Suppose that your sample includes a total of 500 individuals and that, for one of the markers being studied, 498 individuals were homozygous for the common allele whereas 2 others were homozygous for the rare allele. Assuming all 500 individuals are unrelated what is the exact Hardy-Weinberg equilibrium test p-value? Is there evidence for a departure from Hardy-Weinberg equilibrium?

e) If the 500 genotyped individuals are not all unrelated, is the exact Hardy-Weinberg equilibrium test still valid? If so, why? If not, why not?

PROBLEM 2.

The following formula outlines a strategy for calculating the likelihood L observing a set of genotypes $X_1 \ldots X_M$ for a pair of siblings. The formula relies on a nested summation over a set of potential IBD states I.

$$L = \sum_{I_1} \sum_{I_2} \ldots \sum_{I_M} P(I_1) \prod_{i=2}^{M} P(I_i \mid I_{i-1}) \prod_{i=1}^{M} P(X_i \mid I_i)$$

a) What are possible values of I_1?

b) Define $P(I_i|I_{i-1})$ as a function of the recombination fraction θ.

c) Define $P(X_i|I_i)$, allowing for genotyping error.

d) How does the computational complexity of the calculation change as the number of markers M increases? Is that calculation practical when hundreds of markers are examined?

e) Outline a practical approach for evaluating this likelihood when hundreds of markers are involved.

PROBLEM 3.

Your collaborators are interested in studying the contribution of genetic variants to the risk of early onset stroke. They have assembled a large collection of 2000 cases and 2000 controls.

a) Your collaborators plan an initial genomewide association study. They have a genotyping budget of ~$1,000,000 and estimate that genomewide genotyping of each sample will cost ~$500. Should they genotype 1,000 cases and 1,000 controls, spending the entire genotyping budget? Or perhaps they might genotype fewer controls and more cases? What do you recommend?

b) Suppose an initial analysis of the genotype data indicated a genomic control value of 1.03. What does this mean? Should you be concerned?

c) Your collaborators are considering imputing 1000 Genome project genotypes into your samples. Can you name two potential advantages of genotype imputation?

d) After imputation, they notice a genotype dose of 1.30 for marker rs1243 in the individual labeled CONTROL1. What does this mean?

e) How might you evaluate the accuracy of imputed genotypes?

PROBLEM 4.

Your collaborators were so pleased with their initial genomewide association study, that they are now considering a genomewide sequencing study.

a) Describe three potential advantages of a resequencing study compared to a genotyping based genomewide association study.

b) One early step in the analysis of massively parallel sequence data is to accurately place all the reads in relation to each other. A common strategy is generate an index of short words in the reference genome and then use this index to help align each read with respect to the reference genome. Your colleagues are curious about what an appropriate word size might be … 5 letters, 15 letters or 30 letters; why?

c) After alignment, it is standard to calculate the probability of the bases overlapping each particular site for each possible value of the underlying genotype. Please outline this calculation for a homozygous and heterozygous genotype.

d) Suppose that you learned that at typical heterozygous sites, 60% of aligned bases match the reference genome allele and only 40% match an alternate allele. How would this observation change your analysis in c)?

e) Your collaborators are intrigued by "shallow sequencing" approaches, which are much more economical. How do these approaches differ from more standard "deep sequencing" approaches? What are some of the advantages and disadvantages of each method?

PROBLEM 5.

Your collaborators are studying a sample of affected sibling pairs with Crohn's disease. They report that on chromosome 15 they have observed an MLS score of 4.2 and parameter estimates for z_0, z_1, z_2 of 0.30, 0.50, and 0.20 respectively.

a) Define the MLS statistic, with appropriate formulae.

b) What do the estimated z values mean?

c) How would you interpret the observed LOD score?

d) Differences between reported and actual relationships among the individuals being studied can affect the power of a genetic linkage study. How? How might you guard against that possibility?

e) What is the expected kinship coefficient between two siblings? Given a large number of genotypes, could you estimate the kinship coefficient between two individuals without knowledge of marker allele frequencies?