Biostatistics 666
Sample Mid-Term Assessment

Instructions

Please read these instructions 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.

4. I recommend that you **read through all questions** before you attempt to answer any of them.

Problem 1.

Part A

To investigate the role of NOD2 gene mutations in susceptibility to Crohn's disease, a group of researchers collected NOD2 genotype information on 30 patients. In their sample, they found 27 individuals to be homozygous for the wild-type allele, 2 individuals to be heterozygous and 1 individual to be homozygous for the mutant allele.

They decided to first evaluate whether their genotype data fits Hardy-Weinberg equilibrium.

a) Carry out an exact-test to evaluate whether the observed genotypes are in Hardy-Weinberg equilibrium.

b) What are common causes for deviations from Hardy-Weinberg equilibrium? How would you interpret a deviation from HWE in this case?

Part B

Cystic fibrosis is a recessive disease which occurs in individuals with two defective copies of the CFTR gene; individuals with a single defective allele are unaffected. In a sample of 10,000 newborns, 5 affected children were identified.

a) Write out the likelihood of observing a particular number of affected and unaffected children as a function of the frequency of the defective allele, *p*.

b) Outline an E-M algorithm to estimate allele frequencies in the sample. In particular, include the formulas that might be used to update estimated allele frequencies after each iteration.

Problem 2.

Mutations in the G6PD gene, which maps to the X chromosome, are associated with resistance to infection by the malaria parasite. A stretch of 2000bp surrounding the gene was sequenced in a male with was susceptible to malaria and in another male who appeared resistant to malaria.

Assume the effective population size is N = 10,000 individuals, that the mutation rate is 10^{-8} per base-pair per generation and that there is no evidence for recombination in the region.

a) What is the expected time to the most recent common ancestor (MRCA) of the two sequences? Please state any assumptions you made for this calculation.

b) What is the expected number of differences between the two sequences?

c) When the two sequences were compared, 5 differences were identified. What is the probability of observing 5 or more differences between the two sequences? Could you interpret this result as evidence of natural selection at the locus?

d) If your model allowed for recombination within this 2000 bp sequence, how would you expect your answer to a) and b) above to change?

e) In general, how do you expect patterns of genetic variation and linkage disequilibrium to compare between the X chromosome and autosomes? Do you expect to see more (or fewer) variants per base pair in one setting – or do you expect both to be about the same? Do you expect to see the same degree of linkage disequilibrium in both settings – or do you expect one to show greater linkage disequilibrium?

Problem 3.

Genetic variants in the complement factor H (CFH) gene have been associated with susceptibility to age-related macular degeneration, which is a common cause of blindness in the elderly. To study the effect CFH polymorphisms on disease susceptibility, scientists genotyped 3 SNPs (Y402, IVS10 and D936E) in a sample of affected and unaffected individuals.

They observed the following genotype counts among affected individuals:

	Marker		
Y402H	IVS10	D936E	**Counts**
C/C	T/T	A/A	20
C/T	T/G	A/A	10
T/T	G/G	A/G	9
T/T	G/G	G/G	1

And they observed the following genotype counts among unaffected individuals

	Marker		
Y402H	IVS10	D936E	**Counts**
C/C	T/T	A/A	5
C/T	T/G	A/A	15
T/T	G/G	A/G	20
T/T	G/G	G/G	10

a) Describe two methods for evaluating the evidence for association between CFH gene haplotypes and macular degeneration.

b) How could you ensure p-values calculated using the methods you suggested are accurate?

c) Using an E-M algorithm, haplotype frequencies in the combined sample of cases and controls were estimated as $p_{CTA} = 0.417$, $p_{TGA} = 0.300$, $p_{TGG} = 0.283$. All other haplotype frequencies were estimated as zero.

   Calculate D' and r^2 between Y402H and IVS10 and between Y402H and D936E.

d) Do you think it was necessary to genotype all 3 SNPs? Why or why not?

Problem 4.

Assume that you are helping design and analyze a case-control study for a candidate gene.

a) The first step in the analysis is to check that the markers conform to Hardy-Weinberg equilibrium. What is an appropriate way to do this? Why?

b) If multiple markers in the candidate gene are genotyped, outline alternative strategies for estimating haplotype frequencies. Which strategy would you advocate and why?

c) The trait of interest is strongly associated with age. Therefore, when testing for association, investigators would like to account for each individual's age in the analysis. What is an appropriate way to do so? Why?

d) Your collaborators decided to first genotype their case samples and noted that a particular haplotype has estimated frequency of 0.20 in their sample – but only 0.02 in published estimates based on HapMap data. Should they be excited about their finding? What are potential pitfalls?

Problem 5.

Consider a sample of 100 individuals phenotyped for the ABO blood group. Assume that 45, 30, 21 and 4 individuals are found to have blood types A, AB, B and O respectively.

a) What is the likelihood of the observed phenotypes as a function of the three allele frequencies p_A, p_B and p_O?

b) Outline an E-M algorithm to estimate allele frequencies in the sample. In particular, include formulae that might be used to update estimated allele frequencies after each iteration.

c) How would verify that the E-M algorithm identified the maximum likelihood allele frequency estimates?

d) Describe a strategy for calculating a confidence interval for the estimated frequencies.