UNIVERSITY OF CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

Friday 10 June, 2005 1:30 to 3:30

STATISTICAL AND POPULATION GENETICS

Attempt ***THREE*** *questions.*

There are ***FOUR*** *questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

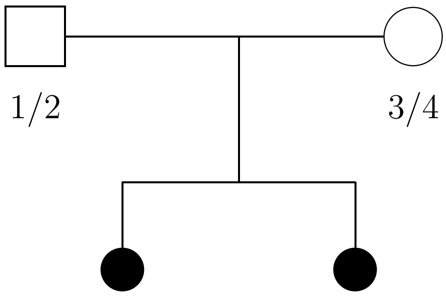
You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 (i) Briefly describe the difference between *parametric* (model-based) and *non-parametric* (model-free) methods of linkage analysis

(ii) Prove that under normal Mendelian inheritance, a pair of siblings share 0, 1 or 2 alleles identical by descent (IBD) at a genetic marker locus with probabilities 0.25, 0.5, 0.25 respectively.

(iii) What is the resulting expected proportion of alleles shared identical by descent by the sibling pair? (Note that the *proportion* of alleles shared corresponds to the *number* of alleles shared divided by 2). How might this be distorted if the marker we are analysing is linked to a disease locus and the siblings are both affected with the disease?

(iv) The figure below shows a pedigree drawing for a family in which the parents are unaffected and the siblings affected with a fully-penetrant recessive genetic disease. The parents are genotyped at a fully informative marker locus as shown.



Denote by L_0, L_1 and L_2, the parametric likelihood function (i.e. the probability of the observed data) for such a family, conditional on the fact that the sibs share 0, 1 or 2 alleles IBD.

Write down expressions for L_0, L_1 and L_2 in terms of θ, the recombination fraction between the disease and the marker locus. Thus show that $\sqrt{L_0 L_2} = L_1$ and $\sqrt{L_0} + \sqrt{L_2} = 0.5$

(v) Suppose that, in a large sample of such affected sib pairs, we observe n_0 pairs sharing 0 alleles IBD, n_1 pairs sharing 1 allele IBD and n_2 pairs sharing 2 alleles IBD. Write down the overall parametric likelihood for the sample in terms of $\sqrt{L_0}$.

(vi) Thus show that the overall parametric likelihood is proportional to

$$s^{2n_0+n_1}(1-s)^{2n_2+n_1}$$

where s is a parameter lying between 0 and 0.5. What does this imply about the relationship between the parametric linkage method and a non-parametric method based on examining the overall proportion of alleles shared IBD?

2 (a) Define the kinship coefficient and derive the kinship between two siblings in an outbred population.

(b) Hence, or otherwise, derive the inbreeding coefficient for a child of a first cousin marriage in an otherwise outbred population.

(c) A disease is caused by a recessive mutation in a single gene and affects one in a million individuals in the (outbred) population. Show that child of a first cousin marriage has 63.4 times the population risk of contracting this disease.

(d) What is the risk to a second child of such a marriage, given that the first child is affected? Does this differ from the sibling recurrence risk in the general population?

(You should state, clearly, any assumptions or approximations you are making.)

3 The number of single nucleotide polymorphisms (SNPs), S, observed in a sample of n chromosomal segments taken from a population is a useful statistic for estimating the mutation rate in that region. This problem explores the properties of S under the standard neutral coalescent. You may assume an infinitely many sites mutation model with mutation parameter θ. The effects of recombination in the segment may be ignored. The time for which the sample has j distinct ancestors is denoted by $T_j, j = 2, 3, \ldots, n$.

(i) Let L denote the total length of the coalescent tree of a sample of size n. Show that the expected value of L is given by

$$\mathbb{E}L = 2\sum_{i=1}^{n-1} \frac{1}{i}.$$

(ii) Give the conditional distribution of S given L, and hence derive the mean of S.

(iii) Suppose we have a prior density $\pi(\theta)$ for θ, and that we observe $S = k$. Denote the coalescence times by $T = (T_2, \ldots, T_n)$, and the posterior density of (θ, T), by $f(\theta, T|S = k)$. Derive a formula for $f(\theta, T|S = k)$.

(iv) Use the result of (iii) to derive a rejection algorithm for simulating observations from $f(\theta, T|S = k)$.

(v) Find the acceptance rate of your algorithm and show how your algorithm can be improved.

(vi) How can you use the output of your algorithm to find an approximate maximum likelihood estimator of θ?

4 This question concerns pairwise linkage disequilibrium observed among markers at many loci in human variation data. Consider a pair of loci, A and B, at each of which there are two possible alleles.

(i) Define D, the measure of linkage disequilibrium (LD) between the loci, and define one other measure of LD.

(ii) Describe the patterns of LD seen across typical human chromosomes.

(iii) How can LD be used for fine-scale mapping of disease genes?

(iv) What is an ancestral recombination graph (ARG), and what is its stochastic structure?

(v) List five biological phenomena that can be explained by use of the ARG, and explain how.

END OF PAPER