UNIVERSITY OF
CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

Friday 4 June, 2004   13:30 to 15:30

Statistical and Population Genetics

Attempt **THREE** *questions.*

There are **four** *questions in total.*

The questions carry equal weight.

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

1 This problem concerns the coalescent process that approximates the evolution of the ancestry of a sample of n chromosomal segments from a population of large but constant size N. Time is measured in units of N generations, and $\theta = 2Nu$ is the mutation parameter, u being the mutation rate per segment per generation. The effects of recombination in the segment may be ignored. Let S be the number of mutations that occur on the coalescent tree of the sample.

(i) Show that the expected value of S is given by

$$\mathbb{E}S = \theta \sum_{i=1}^{n-1} \frac{1}{i}.$$

(ii) Find an expression for the variance of S.

(iii) Using the result of (i), write down an unbiassed estimator $\hat{\theta}$ (say) of θ, and show that it is asymptotically consistent as $n \to \infty$.

(iv) For $j = 2, 3, \ldots, n$, let Y_j be the number of mutations that arise on the coalescent tree while there are j distinct ancestors of the sample. Show that the distribution of Y_j is geometric.

Note: if X has a Poisson distribution with parameter λ, then the probability generating function of X is

$$\mathbb{E}s^X = e^{-\lambda(1-s)}, 0 \leqslant s \leqslant 1.$$

(v) Using (iv) or otherwise, establish that the quantity $\sqrt{\log n}(\hat{\theta} - \theta)$ is asymptotically Normally distributed as $n \to \infty$, and identify the variance.

(vi) What are the practical implications of the result in (v)?

2 One of the major problems in statistical and population genetics is to understand linkage disequilibrium (LD). Write an essay on this topic. You should include brief descriptions of the patterns of LD across chromosome 21, the ancestral recombination graph, its role in understanding LD, and its role in fine-scale mapping.

3 (a) Derive the probability that two first cousins share 0, 1 or 2 alleles identically by descent (IBD)? What is the corresponding kinship coefficient?

(b) What is the inbreeding coefficient for a child of a marriage between first cousins?

(c) For two siblings born to such a marriage, what are the probabilities of 0, 1 or 2 IBD sharing?

(d) Express probabilities of all genotypes in the pedigree extending from two such siblings to their great-grandparents as a graphical model:

(i) as a directed acyclic graph, and

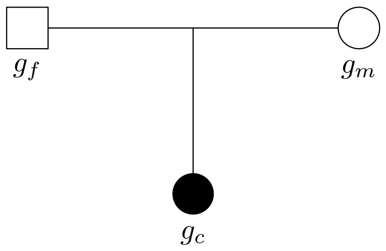
(ii) as a conditional independence graph (CIG).

(e) Indicate links which would have to be added to the CIG to achieve complete *triangulation* of the graph.

(f) After triangulation of the CIG, list all the *cliques* in the graph. Construct a clique junction tree with the *running intersection* property. Explain how this tree helps in probability calculations on the pedigree.

(You may assume, throughout, that there is Hardy–Weinberg equilibrium and random mating in the wider population)

4



The above diagram shows a pedigree drawing for a trio consisting of a father, mother and affected child, with genotypes at a single genetic locus denoted g_f, g_m, g_c respectively.

In a genetic association study, twelve such families are collected with genotypes as tabulated below (where '?' denotes unknown genotypes).

Family	g_f	g_m	g_c
1	2/2	1/1	1/2
2	1/2	1/2	1/1
3	1/2	2/2	1/2
4	1/2	1/2	1/2
5	1/1	2/2	1/2
6	1/2	1/2	1/1
7	1/2	?/?	1/2
8	1/2	1/2	2/2
9	1/2	1/2	1/1
10	1/2	?/?	1/1
11	1/2	1/2	1/2
12	2/2	1/2	1/2

i) For each family, calculate the contribution that it would make to the cells of the following transmission table and thus calculate the values of the counts a, b, c, d in the table.

Transmitted allele	Untransmitted allele	
	1	2
1	a	b
2	c	d

ii) Calculate the value of transmission disequilibrium test (TDT) from this table. Is there any evidence for genetic association? (You may need to know that the percentage points for the upper 5% level are 1.64 for the standard normal distribution and 3.84 for a χ^2 distribution on 1df).

iii) Convert the data in the transmission table to the following table based on unmatched transmissions:

Marker allele	Transmitted	Untransmitted
1	w	y
2	x	z

and use the data in the cells of this table to calculate the haplotype relative risk (GHRR) odds ratio, and test of association.

iv) Prove that for such a trio, the probability of the child's genotypes, given the parents' genotypes and the event (D) that the child is affected with disease $P(g_c|g_f, g_m, D)$ may be written as

$$P(g_c|g_m, g_f, D) = \frac{R_{g_c}}{\sum_{g^* \in G} R_{g^*}}$$

where R_g is the relative risk for genotype g relative to some arbitrary baseline genotype, and the sum in the denominator is over the set G of the four possible offspring genotypes that the parents can produce.

v) Thus prove that the likelihood contribution from the 10 families with both parents genotyped is

$$\frac{R_{1/1}^3 R_{1/2}^4 R_{2/2}}{(R_{1/1} + 2R_{1/2} + R_{2/2})^6 (R_{1/2} + R_{2/2})^2}$$