UNIVERSITY OF
CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

Friday 9 June 2006 9 to 11

STATISTICAL AND POPULATION GENETICS

*Attempt **THREE** questions.*

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

The questions carry equal weight.

STATIONERY REQUIREMENTS	***SPECIAL REQUIREMENTS***
Cover sheet *Treasury Tag* *Script paper*	*None*

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

1 Suppose we have a pedigree consisting of n individuals, with underlying phase-known genotypes at a set of genetic loci $G_1, G_2, \ldots G_n$, and observed phenotypes (disease phenotypes or observed marker genotypes) $X_1, X_2, \ldots X_n$. Denote by $\mathcal{F}$ the set of founders and by $\overline{\mathcal{F}}$ the set of non-founders in the pedigree.

(a) Show that the likelihood (i.e. the probability of the observed phenotype data) for the pedigree, $P(X_1, X_2, \ldots X_n)$, may be written as

$$\sum_{G_1} \cdots \sum_{G_n} \left\{ \prod_i P(X_i|G_i) \prod_{i \in \mathcal{F}} P(G_i) \prod_{i \in \overline{\mathcal{F}}} P(G_i|G_{m(i)}, G_{f(i)}) \right\}$$

where $m(i), f(i)$ denote the parents of individual i.

(b) What factors will the terms $P(X_i|G_i)$ depend on

(i) when X_i is a disease phenotype

(ii) when X_i is an observed marker genotype

and why will these terms often be equal to 0 or 1?

(c) What factors will the terms $P(G_i)$ and $P(G_i|G_{m(i)}, G_{f(i)})$ depend on?

(d) Suppose, rather than studying human pedigrees, we are studing a species in which each individual has a single parent. Consider the simplest pedigree in such a species, consisting of a single parent with a single offspring. Suppose these are phenotyped for a trait governed by a single genetic locus in which only two genotypes are possible. Write out the pedigree likelihood in this case in the same form as the likelihood given above.

(e) Thus show that evaluation of the likelihood in this form results in the computation of 12 multiplications and 3 additions.

(f) How are the number of multiplications and additions altered if we instead use the Elston-Stewart algorithm for evaluation of the likelihood, in which we move the summations as far as possible to the right?

2 (a) Briefly describe the difference between linkage analysis and association analysis.

The table below shows the genotypes at a diallelic locus for a sample of 50 affected individuals (cases), together with the genotypes of their parents.

Mother's genotype	Father's genotype	Case's genotype	Count
1/1	1/1	1/1	1
1/1	1/2	1/1	2
1/1	1/2	1/2	2
1/2	1/1	1/1	2
1/2	1/1	1/2	5
1/2	1/2	1/1	3
1/2	1/2	1/2	7
1/2	1/2	2/2	5
2/2	1/2	1/2	1
2/2	1/2	2/2	2
1/2	2/2	1/2	2
1/2	2/2	2/2	4
2/2	2/2	2/2	6
1/1	2/2	1/2	3
2/2	1/1	1/2	5
			Total: 50

(b) Which rows of this table will *not* contribute to a transmission disequilibrium test (TDT) of association of this locus with disease?

(c) By using the resulting counts in the following transmission table, or otherwise, calculate the value of the TDT for these families. Which allele appears to be the most associated with disease?

Transmitted	Untransmitted allele	
allele	1	2
1	21	20
2	30	29

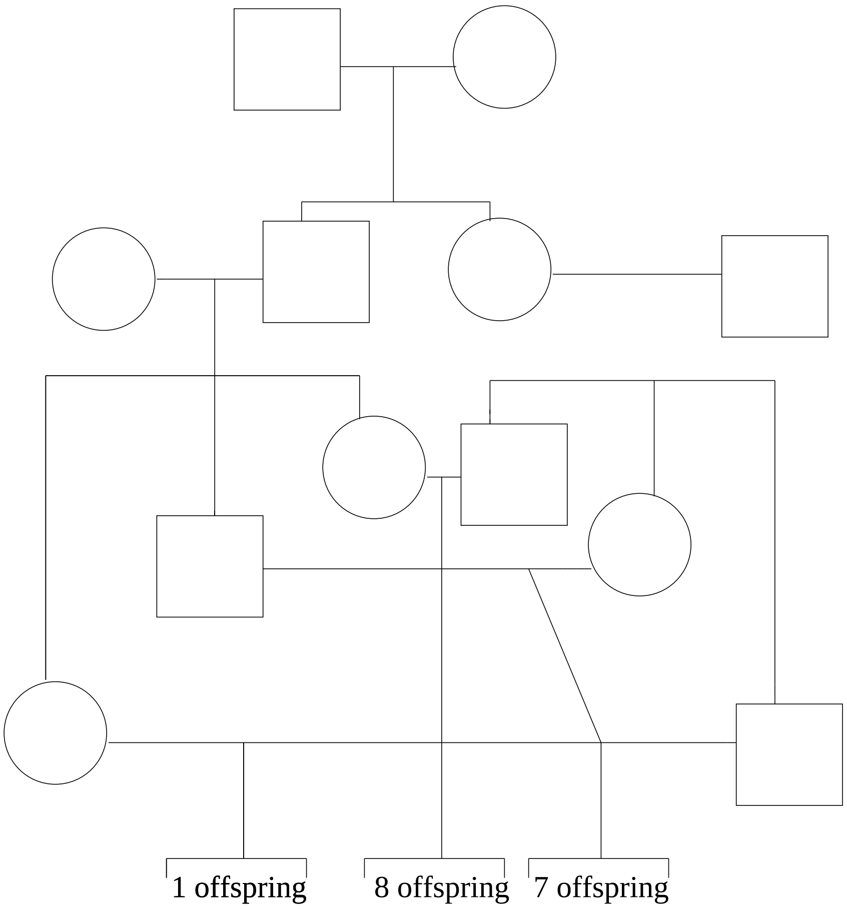
(d) By converting the counts to the form given in the following table, calculate an approximate value for the odds ratio for the effect of allele 2 compared to allele 1.

	Transmitted	Not transmitted
Allele 2		
Allele 1		

(e) The above analysis assumes that the untransmitted alleles can be considered as a control sample to which the transmitted alleles (the case sample) may be compared. What are the three genotype counts (for the genotypes 1/1, 1/2, 2/2 respectively) obtained when considering the untransmitted alleles as control individuals? Do these counts appear to conform to Hardy-Weinberg equilibrium (HWE) proportions?

(f) Briefly describe the advantages and disadvantages of carrying out an association analysis with family data as opposed to using a population-based case/control sample.

3



The 4-generation pedigree above comprises 28 individuals and was studied because 8 of the 16 individuals in the fourth generation exhibited an extremely rare recessive genetic condition. There is considerable inbreeding (three siblings married their first cousins - also siblings), and we can also assume that a defective copy of the gene responsible for the condition is carried by one of the two individuals in the first generation. The condition is so rare that we can safely assume that *no more* than one defective copy entered this pedigree.

In the first group of three questions you should disregard the phenotypic data which led to this pedigree being studied

(a) What is the probability that one of the four copies of a gene in generation 1 is inherited by both siblings in generation 2?

(b) Given this, what is the probability that both partners in all three marriages in generation 3 carry this ancestral copy

(c) Let us single out 1, 4, and 3 individuals respectively from the three sibships in generation 4 (of sizes 1, 7, and 8 respectively). Given that both parents carry the same ancestral copy IBD, what is the probability that the selected individuals in generation 4 each carry *two* copies, and that their remaining 8 siblings do not.

(d) Hence, what is the probability that in these, and only these, subjects in this generation, both maternal and paternal copies of a locus are IBD *and* that they are 2-IBD with each other?

(e) Given the information that these individuals suffer from the condition of interest, and that their siblings do not, what is the posterior probability of this inheritance pattern for the gene which causes the condition? (You may assume that the phenotype is a fully penetrant recessive condition.)

(f) In a linkage study, 10 diallelic markers are typed in a small region. You can safely assume that the region is sufficiently small that no recombination will have occurred in these 4 generations. It was found that all the affected individuals are homozygous for the same haplotype across all 10 markers. Derive an expression for the LOD score for complete linkage ($\theta = 0$) between the gene responsible for the condition and these markers.

(g) Not shown in the pedigree is a further member of generation 3 who married outside the family. She had 3 offspring, of whom only one was affected by the condition. How does the inclusion of these data change the LOD score?

4 The number S of single nucleotide polymorphisms (SNPs) observed in a sample of n chromosomal segments 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 $T_j, j = 2, 3, \ldots, n$.

(a) 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

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

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

(c) The prior density for θ is $\pi(\theta)$, and 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 $f(\theta, T|S = k)$.

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

(e) Find the acceptance rate of your algorithm.

(f) How can you improve the acceptance rate of your algorithm?

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

END OF PAPER