UNIVERSITY OF
CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday 11 June 2001 9 to 12

PAPER 29

BIOSTATISTICS

*Attempt any **FOUR** questions. 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 Survival Data

(a) A continuous survival variable T has integrated hazard function $H_T(t)$. Another survival variable U is related to T by $U = H_T(T)$. Show that U has an exponential(1) distribution.

(b) A survival dataset consists of survival times x_i and visibility indicators v_i ($= 1$ if x_i is an observed failure and $= 0$ if x_i is a censored observation) for each of n individuals.

Assume initially that all individuals are subject to the same survival distribution with survivor function $F_T(t)$. Explain briefly how to obtain the Kaplan–Meier estimate $\hat{F}_T(t)$ of $F_T(t)$. What is the Kaplan–Meier estimate of $H_T(t)$?

(c) A residual u_i is defined by $u_i = \hat{H}_T(x_i)$. What would you expect to see if you constructed a Kaplan–Meier survival curve for the u_i? (Note: u_i is censored when x_i is.)

Suppose the individuals in the dataset can be divided into two groups (for example: Part III and MPhil). State what you would expect to see if you plotted Kaplan–Meier curves for the u_i separately for the two groups

(i) when the survival distribution is the same for all individuals;

and

(ii) when the survival distribution is the same for all individuals in a group, but each group has its own distinct survival distribution.

(d) Another residual y_i is defined by $y_i = v_i - u_i$. Show that the expectation of y_i is approximately zero provided $\hat{H}_T(t)$ is a good estimate of $H_T(t)$.

Discuss briefly how to use the y_i to check whether important explanatory variables have been omitted from $\hat{H}_T(t)$.

2 Survival Data

(a) Write down expressions for

(i) the integrated hazard $H(t)$ in terms of hazard $h(t)$;

(ii) the survivor function $F(t)$ in terms of the integrated hazard;

(iii) the density $f(t)$ in terms of the survivor function.

(b) A survival dataset consists of survival times x_i and visibility indicators v_i ($= 1$ if x_i is an observed failure and $= 0$ if x_i is a censored observation) for each of n individuals. The survival distribution depends on explanatory variables through a parameter vector θ. Derive the log-likelihood for θ in terms of:

(i) f and F;

(ii) h and H.

(c) Suppose the n individuals in fact fall into two groups A and B and it is *not* known which group any particular individual belongs to. The individuals in group A are exposed to hazard $q(t)$ and the individuals in group B are exposed to hazard $q(t) + r(t)$. The probability that an individual belongs to group A is π.

The function $q(t)$ is known for each individual. The functions $r(t)$ and π depend on explanatory variables through parameter vectors ϕ and ψ respectively.

Obtain the log-likelihood for ϕ and ψ in terms of q, r and their integrals.

3 Statistics in Medical Practice

(a) Disturbingly, there are 16 overdose deaths in the 2 weeks after release from prison per annual 10,000 releases of men aged 15-35 years. The prison service has included in the prisoners' pre-release pack a new leaflet about loss of drug tolerance, overdose deaths soon after release from prison, and how to prevent them. It hopes that this intervention will halve overdose deaths in the 2 weeks after release from prison, and proposes to compare overdose deaths among 15,000 releases prior to the intervention, and 15,000 after the intervention.

(i) Show how to estimate the power of the proposed study to detect, with a 5% significance level, a *halving* of the overdose death rate.

(ii) Make one suggestion to improve on the proposed study design.

(b) In 1990, 150,000 adult dairy cattle [over 24 months] were slaughtered for UK consumption at domestic abattoirs where antemortem inspections on the day of slaughter by veterinarians of a random sample of 30,000 cattle detected 15 confirmed BSE cases. The UK herd of adult dairy cattle is about 5 million; early symptoms of BSE are easily missed. Throughout 1990, 14,600 confirmed BSE cases were detected in the field by farmers and their veterinarians. Mean age at clinical onset of BSE is 5 years, and BSE onsets under 30 months are rare, less than 0.5% of all BSE cases.

(i) Compare the 1990 BSE detection rate for adult cattle at antemortem inspection on the day of slaughter, and the day-by-day detection rate by farmers in the national dairy herd of adult cattle.

(ii) Make two suggestions as to why the above detection rates may not be directly comparable.

4 Statistics in Medical Practice

In 1998, hospital X carried out 90 adult heart operations and recorded 9 deaths within 30 days. Suppose that from previous analyses it was estimated that the true underlying mortality rate in such surgery varied between hospitals with a mean of ·05 and a standard deviation of ·02.

(i) Explain how you could choose an appropriate prior density for p, the underlying true rate. How would you then use this, together with the remaining data given above, to obtain a point estimate and an interval estimate for p?

(ii) What qualitative characteristics would these point and interval estimates have, in comparison with the estimate $\hat{p} = 9/90$ that did not make use of the prior information on p?

(iii) Obtain an expression for the Bayes point estimate of p.

(iv) If we now have the corresponding data from all UK hospitals in 1998, under what circumstances should we NOT be willing to assume such hospitals are exchangeable with respect to p?

(v) If we are willing to assume exchangeability for the rates $p_1, \ldots, p_n$ say, from the n hospitals of the UK, what model might be appropriate for the simultaneous estimation of $p_1, \ldots, p_n$?

5 Statistical Genetics

(a) A genetic locus is in Hardy–Weinberg equilibrium in a population, with allele probabilities $\{\pi_i, i = 1, \ldots, n\}$. Assuming the multiplicative model for disease penetrance so that

$$\text{Prob}\,(\text{Disease} \mid \text{Genotype} = i/j) \propto \psi_i \psi_j ,$$

show that the probability of genotype i/j in a subject who has disease is $\pi_i^* \pi_j^*$, where

$$\pi_i^* = \pi_i \psi_i \,/\, \sum_{u=1}^{n} \pi_u \psi_u .$$

What is the implication of this result for the analysis of genetic association studies?

(b) Under the same assumptions, show that the joint probability of genotypes a/b, c/d and a/c in two parents and their offspring, given that the offspring has disease, is

$$\pi_a^* \pi_c^* \pi_b \pi_d$$

— equivalent to drawing a case with genotype a/c and a population "pseudo-control" with genotype b/d in a population–based case–control study.

(c) The probability above does *not* form the basis of the usual analysis of TDT (Transmission Disequilibrium Test) studies, as described in the lecture course. Explain the difference. Why do you think the more usual method is to be preferred?

(question continued on next page)

(d) A TDT study of diallelic locus yielded the following results:

Allele	Transmitted	
Untransmitted	1	2
1	9	24
2	18	49

Express the data as a table of case and pseudo-control *chromosomes*:

	No. chromosomes	
Allele	Case	Control
1		
2		

Show how to calculate the chisquared test for association in this table and compare it with McNemar's test calculated with the same data. Comment on any similarity or dissimilarity you observe.

6 Statistical Genetics

The above pedigree drawings show genotypes of several individuals at 2 genetic loci, locus A and locus B.

(i) In what range is the recombination fraction θ between two loci usually defined?

(ii) What are the haplotypes of individuals 3 and 4 in family 1?

(iii) What is the maximum likelihood estimate of θ, the recombination fraction between loci A and B, using data from family 1 only?

(iv) Write down the likelihood (in terms of θ) for family 2. What is the maximum likelihood estimate of θ using data from family 2 only?

(v) Derive an equation for the maximum likelihood estimate $\hat{\theta}$ using data from both families and show that this equation has a solution in the appropriate range.

(vi) Write down the maximum lod score (in terms of $\hat{\theta}$) for this data. How could one use this to test the hypothesis that locus A and B are linked?