UNIVERSITY OF CAMBRIDGE

**MATHEMATICAL TRIPOS Part III**

Friday 30 May, 2003 9:00 to 12:00

# PAPER 40

## Biostatistics

*Attempt* **FOUR** *questions.*

*There are* **six** *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 Analysis of Survival Data

(a) Write down the density function, survivor function, hazard function and integrated hazard function for an exponential survival distribution with rate parameter $\theta$.

(b) The survival times of a set of $n$ individuals are exponentially distributed with rate parameter $\theta$:

(i) Suppose there is a fixed censoring time $c$: all individual who do not fail before $c$ are censored at $c$. Show that the expected value of the integrated hazard at failure or censoring equals the probability that an individual is observed to fail.

(ii) Suppose instead that the $i$th individual has a fixed censoring time $c_i$: failure is only observed for that individual if it occurs before or at $c_i$. Show that the sum over the $n$ individuals of the expected integrated hazard at failure or censoring is equal to the expected number of individuals that are observed to fail.

(iii) By substituting observed data $(x_1, \ldots, x_n)$ respectively, as times of observed death or censoring, and $d$, as total number of observed deaths, for expectations in the equation derived in part (ii), obtain an estimator for $\theta$.

(iv) Show that the maximum likelihood estimator for $\theta$ is $d/\sum_1^n x_i$. Compare this estimator with the one obtained in part (iii).

(v) Show the second derivative of the log-likelihood at the maximum likelihood estimate of $\theta$ is $-d/\hat{\theta}^2$.

## 2 Analysis of Survival Data

(i) Suppose that the observed failures for a set of $n$ individuals occur at distinct times $a_1 < \ldots < a_j (j \leqslant n)$, and let $r_1, r_2 \ldots r_j$ be the corresponding numbers of individuals at risk at these times. Derive the Nelson–Aalen estimator for the integrated hazard functions.

(ii) Now discuss how to handle ties in the data, i.e. when the failure times for two or more individuals may coincide.

(iii) Use the Kaplan–Meier approach to obtain an estimator for the integrated hazard. Why is there no problem here with tied data?

(iv) Show that for large risk sets the Nelson–Aalen and Kaplan–Meier estimates of integrated hazard are nearly the same.

(v) Show that if there are no censored observations at or after the last observed failure time then the two estimates of the integrated hazard at that time are not nearly the same.

## 3 Statistics in Medical Practice

(i) A tabular CUSUM monitoring scheme is characterised by the updating formula

$$X_t = \max(0, X_{t-1} + W_t), \qquad t = 1, 2, 3, \ldots$$

where $X_t$ represents the value of the process at time $t$, $X_0 = 0$, and $W_t$ is the sample score which is assigned to the $t^{th}$ subgroup of data.

Such a CUSUM procedure is to be used to monitor deaths in an intensive care ward. It is decided to update the process after every $n$ patients treated in the ward. Thus, the available information at time $t$ is the number of observed deaths, denoted $y_t$, for the last $n$ patients treated. Based on national and local data, the probability of death in the intensive care ward is expected to have the value $p_0$.

Assume that the outcomes for different patients are independent.

(1) Assume a death rate of $p_0$ for each patient is taken to represent the null hypothesis. Assume further that alternative hypotheses are defined in terms of an odds ratio parameter which compares the odds of death under the alternative to that under the null hypothesis. For this situation, derive the sample score $W_t$ based on the log likelihood ratio when the odds ratio associated with the alternative hypothesis of interest takes the value $k$.

(2) Briefly explain the difference between the goals of a classical Sequential Probability Ratio Test and a CUSUM plot. Explain how they are effectively combined in a Fast Initial Response CUSUM and when such a monitoring procedure might be used.

(ii) Suppose that the Department of Health intends to publish annual mortality rates for named cardiac surgeons as a league table. These are to be calculated as number of deaths in hospital divided by the number of operations performed. A 95% confidence interval will be provided, thus identifying whether the interval intersects the national mortality rate.

Describe four of the possible statistical objections that could be raised to this practice, and indicate an improvement that might be made in response to each objection.

## 4 Statistics in Medical Practice

One in 200 injectors dies from heroin overdose within two weeks of release from prison. Naloxone, the heroin antidote, if administered to a person who has overdosed, can save life. In Scotland, 8000 injectors on release from prison consent to be randomized, 4000 to carry naloxone and 4000 not to receive it.

(a) Scotland has 20 prisons. Recommend a randomization method to Scottish Prison Service, and explain your choice.

(b) Will the proposed randomization of 8000 injectors have at least 50% power to detect (at 5% significance level) a halving in the overdose death-rate within two weeks of release for injectors randomized to receive naloxone versus controls?

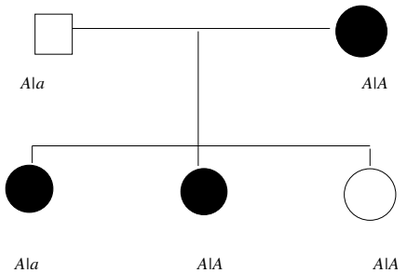
(c) Prison staff point out that, even if randomized to receive naloxone, 20% of released injectors may sell it immediately. Work out the expected number of overdose deaths in the two weeks after release per 1000 injectors randomized to naloxone if 200 immediately sell it.

(d) An international prison-based trial therefore aims to have 80% power to differentiate (at the 5% significance level) between 3 overdose deaths within two weeks of release per 1000 injectors randomized to 'carry naloxone' versus 4.5 overdose deaths within two weeks per 1000 injectors randomized to 'no naloxone'. How many thousands of injectors need to be randomized?

(e) Seventy percent of injectors return to prison within one year. Suggest two questions to ask reincarcerated injectors who had been randomized to naloxone.

## 5 Statistical Genetics

(i) Briefly explain the difference between genetic linkage analysis and genetic association analysis.




(ii) The above pedigree diagram shows a nuclear family in which a fully-penetrant recessive genetic disease is being inherited. Affected individuals are marked in solid black and unaffected individuals in white. Genotypes at a diallelic marker locus are shown beneath each individual.

By inferring the genotypes at the disease locus, write down an expression for the likelihood of the observed data in terms of $\theta$, the recombination fraction between the disease loci. What is the maximum likelihood estimate of $\theta$? Is there evidence for linkage?

(iii) In a population study of the same disease, the following genotype counts among cases and controls were observed.

| Genotype | Cases | Controls |
|---|---|---|
| $a\|a$ | 84 | 16 |
| $A\|a$ | 12 | 48 |
| $A\|A$ | 4 | 36 |

Calculate the expected counts if there is no association between genotype and disease and indicate how these may be used to test the null hypothesis of no association.

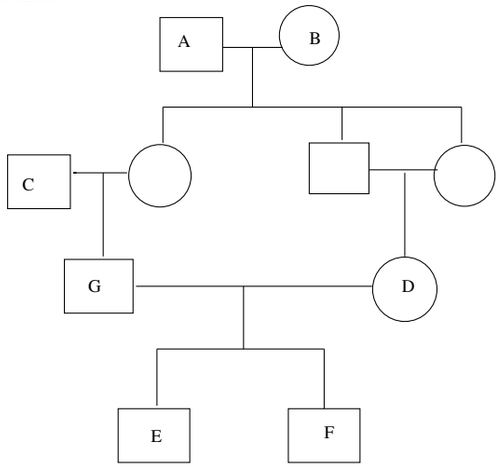
(iv) Write down tables with the observed and expected counts if association is tested at the chromosome (haplotype) level instead of the genotype level, and indicate how these may be used to test the null hypothesis of no association.

(v) Using the observed chromosome table from part (iv), estimate the probabilities that the $A$ and $a$ alleles fall on a disease and a control haplotype respectively. Use these probabilities to obtain a modified expression for the likelihood in part (ii), allowing for linkage disequilibrium between alleles at the disease and marker loci. Show that in this case, the maximum likelihood estimate $\hat{\theta}$ satisfies

$$1.5\hat{\theta}^2 - 2.08\hat{\theta} + 0.54 = 0$$

and write down an expression for the maximum lod score in terms of $\hat{\theta}$.

## 6 Statistical Genetics




(a) Define coefficients of kinship and inbreeding.

(b) In the above pedigree, calculate

(i) the coefficient of inbreeding for individual D,

(ii) the coefficient of inbreeding for individual E,

(iii) the coefficient of kinship for individuals D and E,

(iv) the coefficient of kinship for individuals E and F.

[You may assume kinship and inbreeding coefficients = 0 for founders A, B and C.]

(c) Calculate the mean and variance of the number of alleles shared IBD by a pair of full siblings in the absence of linkage.

(d) In a study of 200 affected sibling pairs at a fully-informative marker locus, the numbers of pairs sharing 0, 1 and 2 IBD were 42, 98 and 60 respectively. Is there evidence of linkage between marker and disease at the 5% significance level?

[Note that the percentage points of the standard normal distribution are 1.64 for the upper 5% level and 1.96 for the upper 2.5% level.]