UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday 8 June 2007 9.00 to 11.00

PAPER 46

BIOSTATISTICS

*Attempt **THREE** questions.*

*There are **FIVE** 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 Statistics in Medical Practice

Watson et al (British Medical Journal, 2005) describe a randomised trial of an intervention providing child safety equipment to prevent injuries to under 5's. They randomised around 1700 families to receive a consultation and free fitted equipment, and 1700 to receive 'usual care'. The primary outcome measure was the rate of injuries requiring medical attendance. For attendance at primary care due to injury the results over a 2-year period were as shown:

	Number of primary care attendances	Child-years at risk	Rate per 1000 child-years
Intervention	220	3595	61.2
Control	172	3888	44.2

(a) What is, approximately, the mean number of children per family?

(b) For a child in a 'usual care' family, roughly what is the chance of attending primary care with an injury in the first 5 years of life?

Watson et al report an estimated rate ratio of 1.37 with 95% confidence interval 1.11 to 1.70, based on a complex multilevel Poisson regression model involving children nested within families.

(c) Suppose individual children had been randomised to intervention and control. Show, without doing any calculations, how an approximate confidence interval for the rate ratio could be obtained. [*Hint: You may assume that if* Y *has a Poisson distribution with mean* m *then, for large* m, $\log(Y)$ *is approximately normal with mean* $\log(m)$ *and variance* $1/m$, *and this variance can be estimated by* $1/Y$]

The resulting approximate 95% confidence interval for the rate ratio is 1.13 to 1.69.

(d) Why does the complex analysis make almost no difference compared with the approximate analysis assuming simple randomisation of children? In what circumstances would you expect there to be a bigger difference?

(e) Give two reasons why the intervention could have increased the risk of attendance at primary care with an injury?

2 Statistics in Medical Practice

Prisoners who have ever injected heroin (ever-injectors) have a high risk of overdose death in the first 4 weeks after release from prison. In Australia, 4-week risk has been estimated as 5 overdose deaths per 1,000 released ever-injectors. In UK prisons, prisoners who have ever injected are randomised to receive on release either Naloxone, the heroin antidote (to be administered to them in the event of overdose), or a control pack which does not contain Naloxone.

(a) How many thousand ever-injector prisoners do the Australians need to randomise for their trial to have 80% power at 5% significance to detect a plausible 30% reduction in overdose deaths in the first 4 weeks after release?

(b) Concern is expressed that one in 5 prisoners randomised to receive Naloxone may have their allocation taken from them by a fellow ex-prisoner released on the same day but randomised to the control group. In the presence of such contamination, how many overdose deaths should you now expect to observe within 4 weeks of release in: (i) 30,000 released ever-injectors randomised to receive Naloxone, (ii) 30,000 released ever-injectors randomised to the control group?

(c) The Australians decided on a 50:50 randomisation of 60,000 ever-injectors but the group randomized to Naloxone actually had 123 overdose deaths whereas the control group had 148 within 4 weeks of release from prison. Provide an approximate 95% confidence interval for the difference in overdose fatalities per 1,000. Are the Australian data consistent with 1.5 fewer overdose deaths per 1,000 randomised to Naloxone?

(d) Suggest how the trial designers could find out about possible contamination between randomized groups.

3 Survival Data Analysis

Let x_i be the censoring time ($v_i = 0$) or event time ($v_i = 1$) of the ith individual in a time-to-event dataset without ties and let $\hat{H}(t)$ be the Nelson-Aalen estimator of the integrated hazard.

Write down a formula for $\hat{H}(t)$ in terms of r_i, the number of individuals in the risk set at x_i, and v_i.

Show that $\sum_i \hat{H}(x_i) = d$ where d is the total number of observed events.

Describe two methods for coping with ties in the data. Which one preserves the property $\sum_i \hat{H}(x_i) = d$?

4 Survival Data Analysis

A time-to-event dataset comprises five individuals:

(i) the event was observed for three of the individuals at times $t = 1, 4$ and 6 respectively.

(ii) the event was not observed for the fourth individual but was known to have occurred strictly after $t = 3$.

(iii) the event was not observed for the fifth individual but was known to have occurred in the interval $2 < t \leqslant 5$.

Write down the empirical likelihood function for the survivor function $F(t)$.

(You may use the notation $F(u-)$ as a shorthand for $\lim_{d \downarrow 0} F(u-d)$.)

By using the fact that F is a decreasing function, show that the maximum empirical likelihood estimator $\hat{F}$ of F satisfies the equations:

$$\begin{aligned}
\hat{F}(1-) &= 1; \\
\hat{F}(4-) &= \hat{F}(3) = \hat{F}(2) = \hat{F}(1) \\
\hat{F}(6-) &= \hat{F}(5) = \hat{F}(4) \\
\hat{F}(6) &= 0
\end{aligned}$$

Applying the corresponding constraints to the likelihood function, show it can be written as

$$(1 - F(2))(F(2) - F(5))^2 F(2)F(5)$$

and maximize it (*hint: take logs*) to obtain $\hat{F}(2) = \frac{4}{5}$ and $\hat{F}(5) = \frac{4}{15}$.

5 Survival Data Analysis

Explain what is meant by *frailty* in survival analysis, illustrating your answer with the proportional frailty model.

The ith individual in a dataset has hazard $U_i\theta$ where the random variables U_i are independently drawn from a Uniform $[\frac{1}{2}, \frac{3}{2}]$ distribution and θ is a constant:

(a) explain what is meant by the *population survivor function* and calculate it;

(b) what would you expect the value of the population hazard function to be for (i) $t = 0$ and (ii) very large t? (Detailed calculations are not required.)

Show that, in general, the population hazard at time t is the expectation of the individual hazards weighted by the individual survivor functions.

[*Hint: The Laplace transform* $\bar{g}(s)$ *of* $g(u) = \mathcal{I}\{a \leqslant u \leqslant b\}$ *is* $\frac{1}{s}e^{-as} - \frac{1}{s}e^{-bs}$.]

END OF PAPER