UNIVERSITY OF
CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

Friday 28 May, 2004 9:00 to 11:00

Biostatistics

Attempt **THREE** *questions.*

There are **five** *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 Survival Data Analysis

Describe what is meant by **censoring**, distinguishing carefully, with examples, between **uninformative** and **informative** censoring. Why is it important to make this distinction?

A researcher is studying student dropout. She obtains her data as follows:

i) whenever a student drops out, the university administrator informs the researcher,

ii) every three months, the researcher writes to the course tutors and asks for a list of students still studying their courses.

The time-to-event variable is the time elapsed from the start of the academic year to the day of drop-out. If the student has not dropped out, then the time-to-event variable is censored at the last day the student was known, through the course-tutor's list, to be studying the course.

The researcher proposes to estimate the survivor function for the first eight months of the academic year.

(a) What difficulties will there be with informative censoring? Illustrate your answer with reference to the following sample data for an academic year starting on 1st October:

student 'A': reported by course tutor to be still studying on 1st April, but for this student the university administrator informs the researcher that 'A' dropped out on 1st May,

students 'B', 'C' and 'D': all these 3 were reported by the course tutor to be studying on 1st April, and there were no reports of these students dropping out.

(b) How would you modify the researcher's design to avoid the censoring difficulties?

2 Survival Data Analysis

(a) (i) Define **survivor function** F, **hazard** h, and **integrated hazard** H for a continuous time-to-event random variable.

(ii) What is the relationship between F and the **density** f?

(iii) Show that $f(t) = h(t)F(t)$.

(iv) What is $F(0)$? What is $H(0)$?

(b) An individual with time-to-event random variable T is subject to a fixed censoring time c. Define $X = min(T, c)$ and consider the expectation $E(H(X))$.

(i) Show that the contribution to $E(H(X))$ from $T > c$ is $F(c)H(c)$.

(ii) Show that the contribution to $E(H(X))$ from $0 \leq T \leq c$ is $1 - F(c) - F(c)H(c)$.

(iii) Hence show that $E(H(X)) = P(T \leq c)$.

(c) Suppose we have a population of n individuals, with

$$X_j = min(T_j, c_j), \text{ for } 1 \leq j \leq n$$

extending the notation of (b) in the natural way. Use part (b) to show that

$$E\Sigma_1^n H(X_j)$$

is the expected number of observed events in this group of n individuals.

3 Survival Data Analysis

Derive the **Nelson-Aalen** estimator for the integrated hazard function of a time-to-event random variable.

Discuss how the **Nelson-Aalen** estimator takes account of (a) censored observations, and (b) tied observations.

A time-to-event dataset consists of n ordered observations, $X_1 < X_2 < \ldots < X_n$, of which d represent event times, and of which the remaining $n-d$ represent censoring times. Show that

$$\Sigma_1^n \hat{H}(X_i) = d$$

where $\hat{H}(t)$ is the Nelson-Aalen estimator of the integrated hazard at time t.

4 Case Studies in Medical Statistics

Assume that convicted heroin injectors are sentenced *either* to 12 weeks in prison *or* to a 1-year drug treatment and testing order (DTTO), which is served in the community (that is, outside prison).

Between sentence date + 1 year and sentence date + 2 years, two-thirds of heroin injectors who had been in jail are expected to be re-convicted. The criminologist Professor UK believes that the alternative of 1-year DTTO will reduce this reconviction rate to one-third. We wish to construct significance tests of size α, and sample sizes to achieve the right power, in helping Professor UK and others to design appropriate studies. Describe the calculations necessary to answer the following questions.

a) How many heroin injectors should be randomised between prison versus DTTO for there to be at least 80% power of detecting whether Professor UK's target has been met, if DTTO's are indeed as effective as Professor UK expects?

b) Practitioners are more sceptical of the effectiveness of DTTO, and consider a reduction in re-conviction rate from two-thirds to one half is a more plausible target. How large would the trial need to be in this case, to achieve the same size α and same power 80%?

c) Deaths from overdose in prison are nil, but deaths from overdose for heroin injectors are 1 per 200 in the first 2 weeks following release from prison, and 1 per 250 in the next 12 weeks following release. Assuming no re-imprisonment within 26 weeks of the original sentence, how many overdose deaths do you expect within 26 weeks of sentence among 10,000 heroin injectors who were sentenced to 12 weeks in prison?

d) Heroin injectors assigned to DTTO are thought to have a 1% overdose mortality in the first 26 weeks following sentence. How many criminal justice systems, each randomising 20,000 heroin injectors equally between 12 weeks' imprisonment (treatment A) and 1-year DTTO (treatment B), would need to collaborate for even 50% power to determine whether the overdose death rate within 26 weeks of release was 12% lower for A than for B?

e) Discuss briefly the ethics of randomising enough injectors to demonstrate lower recidivism (ie re-offending) but not enough injectors to give evidence of raised mortality from overdose.

5 Case Studies in Medical Statistics

The following table shows the mortality rates of 8 hospitals in rank order.

Hospital	Number of operations	Number of deaths	Observed mortality rate
A	10	1	0.10
B	100	12	0.12
C	25	3	0.12
D	100	15	0.15
E	10	2	0.20
F	200	40	0.20
G	20	5	0.25
H	100	30	0.30

Suppose we wish to use a simple graphical display to compare the mortality rate following surgery in hospitals A, . . ., H, and propose plotting the observed mortality rate against the number of operations.

a) Sketch the corresponding graph.

b) Why is this referred to as a 'funnel plot'?

c) Suppose that the Department of Health has stated that the target mortality is 20%. How would you display 'control limits' that indicate a region that you expect would contain 95% of the observed rates if the hospitals were truly 'on target'? (You may assume a normal approximation to the binomial distribution.)

d) List 4 good points about this type of plot.

e) If the 8 hospitals are all truly on target, how many of them do we expect to have mortality rates outside the 95% control limits? Write down an expression for the probability that at least one of the 8 rates will lie outside these limits.

f) Suggest 3 possible ways of dealing with the problem of 'false alarms'.