UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Wednesday, 4 June, 2014 1:30 pm to 4:30 pm

PAPER 32

BIOSTATISTICS

*Attempt no more than **FOUR** questions, with at most **THREE** questions from **Analysis of Survival Data**.*

*There are **SEVEN** 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

Consider a trial in which a new treatment is to be tested. When the new treatment is given to a patient, a normally distributed outcome is observed. Outcomes for patients $1, 2, \ldots$ are i.i.d random variables $Y_1, Y_2, \ldots$, where $Y_i \sim N(\delta, \sigma^2)$. The true value of σ^2 is known.

The null hypothesis $H_0 : \delta = 0$ is to be tested.

(a) From patients $1, \ldots, n$, formulate the Wald test statistic, and derive its distribution i) under H_0 and ii) when $\delta = \delta^*$, where $\delta^* > 0$.

(b) In terms of the standard normal cumulative density function, $\Phi(x)$, derive the formula for the value of n required for a one-sided level-α test with power $1 - \beta$ when $\delta = \delta^*$.

(c) Now consider a two-stage trial with n patients recruited in the first stage and a second set of n patients recruited in the second stage.

Let W_1 label the Wald test statistic of H_0 for the first stage patients ($i = 1, \ldots, n$), and W_2 the Wald test statistic of H_0 for the *combined* set of first and second stage patients ($i = 1, \ldots, 2n$).

Derive the exact joint distribution of (W_1, W_2): i) under H_0 and ii) when $\delta = \delta^*$.

(d) If W_1 is below $f \in (-\infty, \infty)$, a pre-specified finite futility boundary, the trial stops after the first stage without rejecting H_0. If the trial continues to the second stage, H_0 is rejected when $W_2 > \Phi^{-1}(1 - \alpha)$, and is not rejected otherwise.

Show that under this trial design, the total probability of rejecting H_0 when $\delta = 0$ is $< \alpha$.

(e) If the trial continues to the second stage, then it is of interest to estimate δ. It is proposed to use the maximum likelihood estimator $\hat{\delta} = \frac{\sum_{i=1}^{2n} Y_i}{2n}$ to estimate δ.

Explain why this estimator is biased (i.e. $\mathbb{E}(\hat{\delta}|\text{stage 2 occurs}) \neq \delta$) and how this bias will change as δ increases to ∞.

(f) Briefly describe an alternative estimator that uses data from all patients, and is less biased than $\hat{\delta}$. Note that a full derivation of the estimator is not required.

2 Statistics in Medical Practice

Myelodysplastic syndrome (MDS) is a disorder of the blood cells which can lead to leukaemia. It can be classified as mild or severe, and cannot improve over time. MDS can only be cured with hematopoietic stem cell transplantation, but the current clinical policy is to delay transplantation until the disease becomes severe.

A cohort of patients with MDS were observed at a series of clinic visits, beginning at diagnosis and continuing until death. The severity of MDS is recorded at each visit. The following is a sample of clinic visits or death times from three of these patients. The time of death is known exactly, but the disease severity is not recorded when a patient dies.

Patient number	Months after diagnosis	State (1=mild, 2=severe, 3=death)
1	0.0	1
1	8.5	1
2	0.0	1
2	26.3	1
4	0.0	2
4	12.6	2
4	47.2	3

It is proposed to fit a continuous-time Markov multi-state model to the full data to represent the process of MDS progression along with death from any cause.

(a) Draw a diagram of the states and allowed transitions in the model, including symbols for the corresponding transition intensities, and write down the transition intensity matrix (with as few unknown expressions as possible).

(b) Suppose that the researchers

- expect patients in this cohort with mild MDS to spend about 8 years before progression or death,
- believe progression or death from mild MDS is equally likely, and
- expect patients with severe MDS to survive about 3 years.

Derive the corresponding elements of the (monthly) intensity matrix from (a).

(c) Write down a formula for the contribution of the three patients above to the likelihood as a function of the transition intensity matrix. Define any terms that you use and state all assumptions that are being made in this model.

The approximate intensities obtained in (b) were used as initial values in a numerical algorithm to compute the maximum likelihood estimates under this model. The maximum likelihood estimates for the monthly transition intensities are as follows, with the corresponding 95% confidence intervals in brackets.

Mild – Mild	-0.0115	(-0.0130,-0.0102)
Mild – Severe	0.0072	(0.0061, 0.0085)
Mild – Death	0.0043	(0.0035, 0.0054)
Severe – Severe	-0.0318	(-0.0352,-0.0287)
Severe – Death	0.0318	(0.0287, 0.0352)

(d) What is the expected time spent before moving to another state, for a person with mild MDS, and severe MDS, respectively, with 95% confidence intervals? Hence calculate the expected survival time for a person in mild MDS (no confidence interval required).

The following table gives the maximum likelihood estimates (with 95% confidence intervals in brackets) for a further model in which each log transition intensity is a linear function of two binary covariates: indicators for whether the previous observation was made in the period from 0–3 months after a hematopoietic stem cell transplant, and >3 months after transplant, respectively.

	Transition intensity (before transplant)	Hazard ratio after transplant 0-3 months	>3 months
Mild – Mild	-0.0119 (-0.0136,-0.010)		
Mild – Severe	0.0103 (0.0087, 0.012)	0.52 (0.11, 2.4)	0.02 (0.0031,0.15)
Mild – Death	0.0016 (0.00009, 0.003)	43.92 (21.07,91.5)	3.54 (1.73,7.25)
Severe – Severe	-0.0380 (-0.044,-0.033)		
Severe – Death	0.0380 (0.033, 0.044)	2.37 (1.80, 3.1)	0.57 (0.46,0.72)

(e) Interpret the reported hazard ratios for the covariate effects. Suggest plausible explanations why particular ones are greater (or less) than 1, and why some are bigger than others. How might these results justify the transplantation policy stated in the introduction to this question?

3 Statistics in Medical Practice

(a) When analysing a dataset that contains missing values, it is common to assume that the data are missing at random (MAR). Write down an equation that defines the missing at random assumption. Briefly define the notation used in your equation.

(b) A sample of 102 individuals newly diagnosed with a particular medical condition are recruited into a cohort study and followed up over two years. At the end of each of the two years, the individuals are asked whether they have used a prescription drug in the past 12 months. Unfortunately, some patients drop out of the cohort study during the second year, and once a patient has dropped out there is no further opportunity to question him or her. The following table gives the numbers of patients with each possible pattern of data ('–' denotes that data are missing because the patient has dropped out).

Pattern of data		Number of
Used a prescription drug in year 1	Used a prescription drug in year 2	individuals with this pattern
no	no	30
no	yes	7
yes	no	10
yes	yes	18
no	–	21
yes	–	16

(i) You wish you know whether these data can be assumed to be missing at random, but the doctor managing this cohort study does not know what 'missing at random' means. What would be a sensible way to ask this doctor whether she thinks the missing at random assumption is reasonable for these data?

(ii) Assume that the data are missing at random and estimate the probability that a patient uses a prescription drug in year 2.

(iii) Now suppose that you can assume that the data are missing completely at random. Using this assumption, calculate a more efficient estimate of the probability that a patient uses a prescription drug in year 2. Give a brief explanation for why this estimator is more efficient than the one you calculated in part (b).

(c) Suppose that you have a sample of N individuals and you wish to investigate how blood pressure depends on age. Let Y_i denote the ith individual's blood pressure and let X_i denote his or her age. You specify a linear regression model with blood pressure as the outcome and age as the covariate. Unfortunately, although blood pressure is observed for all N individuals, age is missing for some of them. Your colleague tells you that if you assume that these data are missing at random, then the missingness pattern is ignorable. Explain what this statement means and provide a formal proof that it is true. [You may assume that the parameters of the model for the data and the parameters of the model for the missingness pattern given the data are distinct.]

4 Analysis of Survival Data

What is meant by the *intensity* of a counting process? Derive the *Nelson–Aalen* estimator of the integrated hazard. [You may assume there are no ties in the data.]

A time-to-event dataset is made up of the following observations:

3, 4+, 5+, 6, 9,13

where a '+' indicates a censored observation.

Calculate the Nelson–Aalen estimate of the integrated hazard at each of the timepoints. Show that the sum of the estimates of the six integrated hazards equals the number of observed events.

Let there, in general, be n individuals and denote the observation time (event or censored) of the ith individual be x_i. Continuing to assume no ties, prove algebraically that $\sum_{i=1}^{n} \hat{H}(x_i)$, where $\hat{H}(t)$ is the Nelson–Aalen estimate of the integrated hazard, is equal to d, the number of observed events.

Suppose now that it is assumed the hazard function is a constant θ. If $\hat{\theta}$ is an estimate of θ, what now is $\hat{H}(x_i)$? Taking $\sum_{i=1}^{n} \hat{H}(x_i) = d$ to be a desirable general property of estimators of integrated hazards, derive an estimator for θ and calculate its value for the time-to-event dataset. Do you think it is a good estimate?

5 Analysis of Survival Data

A continuous time-to-event random variable T has integrated hazard function $H(t)$. Show that the time-to-event random variable U defined by $U = H(T)$ has an `exponential`(1) distribution whatever the form of H. [You may assume that H has an inverse.]

A time-to-event dataset is made up of observations (x_i, v_i), $i = 1, \ldots, n$ where if $v_i = 0$ then x_i represents a censored observation and if $v_i = 1$ then x_i represents an observed event. A model has been fitted to the data and the integrated hazard has been estimated. A new time-to-event dataset (y_i, v_i) is constructed where $y_i = \hat{H}_i(x_i)$. Explain briefly how you can use this new dataset to investigate the adequacy of the model.

If the model is correct, and there are no censored observations, what would you expect the mean of the y_i to be approximately equal to? If there are censored observations, how would you adjust the y_i to give, over all the observations, an approximately known expected mean? Justify your answer in general terms.

Verify that your proposed adjustment is reasonable in the following special case. The time-to-event variable U has an `exponential`(1) distribution. The time-to-censoring variable C has, independently of U, a probability π of being equal to c and a probability $1 - \pi$ of being equal to ∞. Define a new time-to event variable U^* which equals U if the event is observed and which equals $C + k$ if the event is censored, that is: $U^* = \min(U, C) + k\mathbb{I}[C < U]$ where k is a constant and $\mathbb{I}$ is the indicator function. Show that k can be chosen such that $\mathbb{E}U^* = 1$. Comment on the implications of k's lack of dependence on c.

6 Analysis of Survival Data

Explain what is meant by a *Schoenfeld function* and by a *Schoenfeld residual*. How are the Schoenfeld residuals calculated after a proportional hazards model has been fitted? Outline briefly how Schoenfeld residuals can be used to detect variation with time of the dependence of hazard on explanatory variables.

A time-to-event dataset has been generated by a proportional hazards process such that the hazard for the ith individual ($i = 1, \ldots, n$ with $n \geqslant 3$) is given by $h^i(t) = \exp(\beta_0 z^i) h_0(t)$ where β_0 is a constant, $z^i \in \{0, 1\}$, and $h_0(t)$ is the baseline hazard. At time $t = \xi$ there are precisely three individuals in the risk set, with distinct observed event times. For two of those individuals $z^i = 0$ and for the third $z^i = 1$.

(a) Given that there is an event at $t = \xi$, calculate the probability – conditional on the history of the time-to-event process up to just before $t = \xi$ – that it is an individual with $z^i = 0$ who has the event. Write down the correponding Schoenfeld function calculated at $\beta = \beta_0$. Similarly, write down the conditional probability and the corresponding Schoenfeld function calculated at β_0 when the individual with $z^i = 1$ has the event. Show that the conditional expectation of the Schoenfeld function calculated at β_0 is zero.

(b) Suppose that it is in fact the individual with $z^i = 1$ who has an event at $t = \xi$. Write down the Schoenfeld function $s(\beta)$ for variable z at time ξ. Calculate $s(-\infty)$, $s(0)$ and $s(\infty)$, interpreting your answers.

7 Analysis of Survival Data

What, in the context of *competing risks*, is meant by a *cause-specific hazard*? What is meant by a *cumulative risk function*? Obtain an expression, in terms of the cause-specific hazards, for the cumulative risk function of a particular event in the presence of competing events.

(a) Patients with a particular disease undergo surgery at time $t = 0$. After surgery, patients are exposed to the risk of death from two sources:

A: due to the surgery – the cause-specific hazard equals θ_A for $t \leqslant \tau$ and equals zero for $t > \tau$.

B: due to the disease (despite surgery) – the cause-specific hazard equals θ_B for all t.

Obtain the cumulative risk function for death due to the disease, evaluating any integrals. What is the probability that an individual dies as a consequence of surgery? Verify that all individuals ultimately die from one cause or another.

(b) Using the following data, calculate a non-parametric estimate of the cumulative risk function for event B at time $t = a_{k+3}$ in the presence of a competing event A:

(i) the estimate of the cumulative risk function for event B at $t = a_k$ is 0.3;

(ii) the estimate of the survivor function for the composite event 'A or B' at $t = a_k$ is 0.4;

(iii) there are ten individuals in the risk set just after $t = a_k$;

(iv) no individuals have events or are censored in the interval $a_k < t < a_{k+1}$;

(v) an individual has event B at $t = a_{k+1}$;

(vi) no individuals have events or are censored in the interval $a_{k+1} < t < a_{k+2}$;

(vii) an individual has event A at $t = a_{k+2}$;

(viii) no individuals have events or are censored in the interval $a_{k+2} < t < a_{k+3}$;

(ix) an individual has event B at $t = a_{k+3}$, and another individual is censored at that time.

END OF PAPER