UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 6 June, 2016 9:00 am to 12:00 pm

PAPER 207

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

Let there be three treatments for a condition that are to be tested in a clinical trial: treatment 0 (the control treatment), treatment 1 and treatment 2.

Let the number of patients allocated to arm i be n_i. The jth patient allocated to arm i has treatment response Y_{ij}, which is normally distributed with mean μ_i and variance σ_i^2. The value of μ_i is unknown but σ_i^2 is known. Two null hypotheses are to be tested in the trial:

$$H_0^{(1)} : \delta_1 = \mu_1 - \mu_0 \leqslant 0$$
$$H_0^{(2)} : \delta_2 = \mu_2 - \mu_0 \leqslant 0$$

(a) Write down the Wald test statistics W_1 and W_2 for testing $H_0^{(1)}$ and $H_0^{(2)}$ respectively.

(b) Derive the joint distribution of (W_1, W_2).

(c) Let $H_0^{(i)}$ be rejected if $W_i > c, i = 1, 2$. Write down the probability of rejecting at least one of $H_0^{(1)}$ or $H_0^{(2)}$ in terms of a multidimensional integral of a function you should specify.

(d) By showing the probability of rejecting at least one null hypothesis is increasing in both δ_1 and δ_2 or otherwise, show that the maximum chance of making a type I error in this trial is when $\delta_1 = \delta_2 = 0$.

(e) Describe *either* a group-sequential design *or* a response-adaptive randomised rule. In your answer you should describe briefly what the relevant procedure is together with two potential benefits and two drawbacks of using it for this trial.

2 Statistics in Medical Practice

A population of people at risk of developing a particular kind of cancer are given a screening test every t years. The test may reveal that they have pre-clinical cancer, which can be treated before it progresses to clinical cancer.

There are data consisting of the t-yearly screening results, and, if clinical cancer is detected at times other than screen, the times of these diagnoses. It is proposed to use a continuous-time Markov model to represent the onset and progression of the cancer in the absence of treatment. Assume that all cases of this cancer have a pre-clinical phase.

(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) Write down or derive expressions for:

(1) the probability that a patient who has just been screened as cancer-free develops pre-clinical cancer before their next screen;

(2) the probability that a patient, who has developed pre-clinical cancer at some time $u < t$, goes on to develop clinical cancer before their next screen at time t;

(3) the probability that a patient just screened as cancer-free develops *clinical* cancer before their next screen. (It may be assumed that the onset rate of pre-clinical cancer is different from the progression rate).

Hence write out the transition probability matrix for the Markov model.

(c) The Markov model was fitted to data, giving a maximum likelihood estimate for the transition intensity from the disease-free to the pre-clinical state of $q_{12} = 0.005$. Given a population of 1000 people, how many of them are expected to get preclinical cancer over a period of 10 years?

(d) Suppose a portion of the population is known to have a specific genetic risk factor for this kind of cancer. It is also known that risk of cancer onset changes smoothly with increasing age for everybody.

- Write down a formula which may be used to extend the model in (c) to relate a patient's age and presence of the genetic predictor to their risk of onset, defining all terms used.
- From such an extended model, suppose we have estimated the rate of pre-clinical cancer onset for a 50 year old person without the risk factor to be q, and the hazard ratios for the genetic risk factor and one year of age are α_1 and α_2 respectively. What is the estimated rate of pre-clinical cancer onset for a 60 year old with the risk factor?

(e) Under the model in (d), explain why we cannot obtain a closed form for the transition probability matrix $P(t_1, t_2)$ between a pair of times t_1, t_2. Given data from one patient consisting of observed states at a discrete set of times $\{t_i : i = 1, 2, \ldots\}$, what approximation to the covariate values might we use to construct a closed form likelihood for this model?

(f) Assume now that a proportion of people with harmless tumours are expected to be wrongly diagnosed with pre-clinical cancer, but otherwise the screening test is accurate. Given data from one patient consisting of two negative screens at years 0 and 2 and a positive screen at year 4, obtain the likelihood in as simple a form as possible, as a function of specific parameters, defining any new parameters. No covariates need to be included.

3 Statistics in Medical Practice

Assume we have a closed population of size $N+1$ (i.e. with no births, deaths, immigration or emigration in the time period we are considering), for example the population of a boarding school. Suppose at time 0, one of the $N+1$ students is infected with measles. Once infected, assume an individual is infectious for a mean period of γ^{-1} days before becoming immune, and hence recovered and "removed" from the susceptible population. Assume also that students in the school are homogeneously mixing.

(a) Draw a state diagram describing the dynamics of measles as described above, using and defining standard notation for the states and transition rates.

(b) Write down a deterministic system of equations describing this compartmental model, including also any initial conditions and the support of any transition rates.

(c) From the system of equations and the initial conditions, derive expressions, in terms of N and the number of recovered students at time t, for: (i) the number of susceptible students; (ii) the number of infectious students.

(d) Show that the threshold result $N > \gamma/\beta$ holds if the epidemic takes off, where β is the effective contact rate and γ is the recovery rate.

Assume now a chain-binomial discrete-time stochastic representation of the above system, where the system is specified as

$$\begin{aligned} S(t+\delta t) &= S(t) - B(t) \\ I(t+\delta t) &= I(t) + B(t) - C(t) \\ R(t+\delta t) &= R(t) + C(t) \end{aligned}$$

where $B(t)$ and $C(t)$ are assumed to be independent Binomial random variables representing the numbers of new infections and new recoveries respectively occurring in $[t, t+\delta t)$:

$$\begin{aligned} B(t) &\sim \text{Bin}(S(t), \beta I(t)\delta t) \\ C(t) &\sim \text{Bin}(I(t), 1-\exp(-\gamma\delta t)) \end{aligned}$$

where the time units $\delta t = 1$ day are such that $1-\exp(-\beta I(t)\delta t) \approx \beta I(t)\delta t$.

(e) Show that the probability of extinction at time $t=1$ can be expressed as

$$(1-\beta)^N(1-\exp(-\gamma)).$$

Alternatively, the epidemic could be formulated to be in continuous-time and stochastic. To simulate the epidemic requires the use of the Gillespie algorithm.

(f) Write down the system of stochastic equations and initial conditions.

(g) What is the distribution for the time until the first event (i.e. the first change of state)?

(h) What is the probability that this first event leads to extinction of the epidemic?

(i) Show that, for $N \geqslant 2$, the probability that there is more than one event in the first interval $[0, 1)$ is

$$\frac{\beta N}{\beta^*}\left(1 - e^{-\beta^*}\right) - \frac{\beta N e^{-\beta^*}}{\beta^* - 2\beta}\left(1 - e^{-(\beta^* - 2\beta)}\right)$$

where $\beta^* = \beta N + \gamma$.

4 Analysis of Survival Data

Outline how to construct the *Kaplan-Meier* estimator of the survivor function $F(t)$ in terms of r_j, the number of individuals at risk at time a_j, and d_j, the number of individuals with an event at that time, for a suitable set of times a_j.

A time-to-event dataset comprises seven individuals. Six of the seven individuals have event times given by $t_1, \ldots t_j, \ldots t_6$ with $t_j < t_{j+1}$. The seventh individual is censored at time c_7 where $t_3 < c_7 < t_4$.

(a) Calculate the Kaplan-Meier estimate of $F(t_5)$ (call this estimate $\hat{F}_5^*$).

Given that an individual's event time is $\geqslant t_4$, what are the estimated conditional probabilities $\hat{p}_4^*, \hat{p}_5^*$ and $\hat{p}_6^*$ of the individual having an event at t_4, t_5 and t_6 respectively?

(b) Assume that the seventh individual's event time is now known to be equal to the fourth actually observed event time (that is: $c_7 = t_4$). Calculate the Kaplan-Meier estimate of $F(t_5)$ under this assumption (call this estimate $\hat{F}_5^0$).

Calculate, also under this assumption, the estimated conditional probabilities $\hat{p}_4^0$, $\hat{p}_5^0$ and $\hat{p}_6^0$ of an individual having an event at t_4, t_5 and t_6 respectively, conditional on the event time being $\geqslant t_4$.

(c) Re-calculate the Kaplan-Meier estimate of $F(t_5)$, now splitting the seventh individual's event over times t_4, t_5, t_6 in the proportions $\hat{p}_4^0, \hat{p}_5^0, \hat{p}_6^0$ (call this estimate $\hat{F}_5^1$). *Hint*: Justify that the seventh individual's contribution to d_j is $\hat{p}_j^0$ and to r_j is $\sum_{j':j' \geqslant j} \hat{p}_{j'}^0$ for $j \geqslant 4$.

Comment on the numerical values of $\hat{F}_5^0$, $\hat{F}_5^1$ and $\hat{F}_5^*$.

(d) Show that the estimate of $F(t_5)$ obtained by splitting the contribution of the seventh individual's event over the times t_4, t_5, t_6 in the proportions $\hat{p}_4^*, \hat{p}_5^*, \hat{p}_6^*$ is equal to $\hat{F}_5^*$, and comment.

5 Analysis of Survival Data

Show that, if T is a time-to-event variable with integrated hazard function H, the variable $U = \mathrm{H}(T)$ has an `exponential`(1) distribution. (You may assume H has an inverse.)

(a) A time-to-event dataset $\{(x_i, v_i)\colon i = 1, \ldots, n\}$ comprises n individuals: x_i being either the time of the observed event ($v_i = 1$) or the time of censoring ($v_i = 0$) for the ith individual. Let $\mathrm{H}(t)$ be the common integrated hazard and $\hat{\mathrm{H}}(t)$ be an estimated integrated hazard obtained from a model. What would you expect the Kaplan-Meier plot (with log-transformed vertical axis) of the time-to-event dataset $\{(\hat{\mathrm{H}}(x_i), v_i)\colon i = 1, \ldots, n\}$ to look like, assuming the model is a good fit to the data?

(b) Let $Y = \min(\mathrm{H}(T), \mathrm{H}(C))$ where C is a time-to-censoring variable. What is known about the expectation of Y? How would you modify the definition of Y to give a variable with known expectation?

Describe how the modified Y can be used to explore the contribution of explanatory variables to a time-to-event model.

6 Analysis of Survival Data

(a) Define the *survivor* function of a continuous time-to-event variable. How is the density function related to the survivor function ?

What is meant by a *hazard* function? Write down an expression for the hazard function in terms of the density and the survivor functions.

What condition is imposed on the density if the event is certain to happen at some time (in the absence of censoring). What, in that case, is the limit of the integral from 0 to t of the hazard function as $t \to \infty$?

(b) A time-to-death study enters individuals at a uniform rate between calendar times τ_a and τ_b. The study closes at calendar time τ_c ($\tau_a < \tau_b < \tau_c$) when all individuals who have not died are censored. There is no other source of censoring. State, with justification, whether or not this is an example of informative censoring.

Let the time-to-death (from study entry) variable be T and the time-to-censoring (from study entry) variable be C.

(i) Obtain the density, survivor, hazard and integrated hazard functions for C in terms of t, the time since the individual's entry into the study. Interpret the behaviour of the hazard and integrated hazards near $t = \tau_c - \tau_a$.

(ii) Obtain an expression for the probability of an individual either dying or being censored by time t since study entry ($0 \leqslant t < \tau_c - \tau_a$) in terms of τ_a, τ_b, τ_c and the survivor function for T.

7 Analysis of Survival Data

The ith individual of a population of n individuals is subject to a continuous time-to-event process with density $\mathrm{f}_i(t)$ and hazard $\mathrm{h}_i(t)$.

(a) Show that the population density $\bar{\mathrm{f}}(t)$ is the mean of the individual densities. Why, in general, cannot the population hazard $\bar{\mathrm{h}}(t)$ be the mean of the individual hazards? Obtain an expression for the population hazard as a weighted mean of the individual hazards.

(b) Estimation of integrated hazard is often based on the equation

$$\mathbb{E}\{\mathrm{d}N_i(t)|\mathcal{H}_{t-}\} = Y_i(t)\mathrm{h}_i(t)\,\mathrm{d}t$$

where $N_i(t)$ ($\in\{0,1\}$) is the indicator for the ith individual having an observed event before or at t, $Y_i(t)$ is the indicator for the ith individual being at risk at t and the expectation is conditional on the history $\mathcal{H}_{t-}$ up to but not including t.

(i) Interpret this equation. What is meant by the 'history' in this context?

(ii) In the absence of censoring, what is $\mathbb{E}Y_i(t)$ (unconditionally)?

(iii) If the time-to-censoring variable has common survivor function $G(t)$, and is independent of the time-to-event variable, what is $\mathbb{E}Y_i(t)$ (unconditionally)?

(iv) What then is the unconditional expectation of $\mathrm{d}N_i(t)$?

(c) Suppose the n individuals are subject to a common integrated hazard: $\mathrm{H}_i(t) = \mathrm{H}_0(t)$. Derive $\hat{\mathrm{H}}_{NA}(t)$, the Nelson-Aalen estimator of $\mathrm{H}_0(t)$. (Keep your answer in the form of an integral.)

(d) Suppose now that the $\mathrm{H}_i(t)$ did depend on i but all individuals are subject to an independent time-to-censoring process with common survivor function $G(t)$. Give an informal derivation of the unconditional expectation of $\hat{\mathrm{H}}_{NA}(t)$ in terms of the $\mathrm{H}_i(t)$. (You may use the approximation $\mathbb{E}U/\mathbb{E}V$ for $\mathbb{E}\{U/V\}$.)

Comment on the dependence of your answer on $G(t)$. Is $\hat{\mathrm{H}}_{NA}(t)$ an unbiased estimator of the population integrated hazard $\bar{\mathrm{H}}(\mathrm{t})$?

END OF PAPER