UNIVERSITY OF
CAMBRIDGE

MAT3, MAMA

MATHEMATICAL TRIPOS Part III

Thursday, 30 May, 2019 1:30 pm to 4:30 pm

PAPER 207

STATISTICS IN MEDICINE

*Attempt no more than **FOUR** questions.*

*There are **SIX** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper
Rough 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

Suppose we observe retrospective data on an epidemic of influenza. These data cover a single season (specifically winter 2017-2018) of a particular strain of influenza in a small fixed population of size N, with no births, deaths or migration during the season. Let I denote the number of individuals who have experienced influenza infection by the end of the season. Let Y_G denote the number, out of the I infected individuals, who visited their GP about their influenza infection. Let ρ_G denote the probability of visiting a GP conditional on influenza infection. We are interested in estimating the probability of infection during the 2017-2018 season, λ. Assume that each individual can only be infected by the particular strain once during the season.

One possible simple Bayesian model relating Y_G to λ is:

$$\begin{aligned} Y_G \mid (N, \lambda, I, \rho_G) &\sim \text{Binomial}(I, \rho_G) \\ I \mid (N, \lambda) &\sim \text{Binomial}(N, \lambda) \\ \rho_G &\sim \text{Beta}(a, b) \\ \lambda &\sim \text{Beta}(1, 1) \end{aligned}$$

where a and b have fixed values, and the parameters λ and ρ_G are assumed *a priori* to be independent of each other and of N.

(a) Draw the Directed Acyclic Graph (DAG) corresponding to this model.

(b) In previous influenza seasons, the estimated conditional probability of visiting a GP given influenza infection varied from season to season, but averaged 0.1 and had a standard deviation of 0.02985. Calculate which values of a and b should be chosen so that the Beta(a, b) distribution reflects this prior knowledge about ρ_G. Note that the mean and variance of a Beta(a, b) distribution are $a/(a+b)$ and $\frac{ab}{(a+b)^2(a+b+1)}$ respectively. Note also that 0.02985 can be expressed as $\left(\frac{9}{10{,}100}\right)^{1/2}$.

(c) Show either by summing over possible values of I or otherwise that $Y_G \mid (N, \rho_G, \lambda) \sim$ Binomial$(N, \rho_G\lambda)$. Derive therefore an expression for the likelihood of the data y_G given the basic parameters ρ_G and λ.

(d) Write down an expression for the posterior distribution of the parameters ρ_G and λ given the data y_G.

We now consider synthesising further evidence that informs the probability of infection λ. Assume that a proportion π_1 of the population is immune to the 2017-2018 strain before the start of the season, corresponding to the prevalence of previous influenza infection. Assume these individuals remain immune throughout the season so that they cannot be infected during the season. Let π_2 denote the prevalence of infection with the 2017-2018 strain at the end of the 2017-2018 season. You can then assume that π_2 equals the sum of the probability of being immune at the start of the season and the probability of infection λ during the season.

[QUESTION CONTINUES ON THE NEXT PAGE]

Suppose that we conduct two independent prevalence surveys, one just before the start of the 2017-2018 influenza season and a second at the end of the season. Sampled individuals have their blood tested for evidence of influenza infection in the past. Assuming that both samples are representative of the whole population of size N, π_1 and π_2 can be estimated from the observed number testing positive for the strain Y_{Si} out of n_{Si} samples tested, for $i = 1, 2$.

(e) Write down an appropriate model (specifying both distributional and functional assumptions) for how the survey data y_{S1} and y_{S2} relate to the probability of infection λ. Assume that your prior distribution for (π_1, λ) should be uniform on the support of (π_1, λ).

(f) Derive the marginal prior distributions of π_1 and π_2.

(g) Extend your DAG from part (a) to include the survey data.

(h) Synthesising both the survey and the GP data, write down an expression for the posterior distribution of your parameters given the observations y_G, y_{S1} and y_{S2}.

2 Statistics in Medical Practice

1. What is the objective of a mediation analysis? Provide a definition of a natural direct effect and a natural indirect effect in words, and using counterfactual notation for a binary exposure variable X, a binary mediator M, and an outcome Y.

2. Please read the shortened paper provided as supplementary material. If we are prepared to assume the assumptions for the mediation analysis are satisfied, what does this analysis tell us about the causal role of smoking intensity as a risk factor for lung cancer?

3. What evidence does the interaction analysis provide about the causal role of smoking as a risk factor for lung cancer?

4. Provide and explain some reasons why the mediation analysis could give misleading results.

5. How could the authors improve their existing analysis to strengthen evidence relating to the causal role of smoking as a risk factor for lung cancer?

[SUPPLEMENTARY MATERIAL ON THE NEXT PAGE]

Question 2 Supplementary Material: Genetic Variants, Smoking and Lung Cancer: An Assessment of Mediation and Interaction

Genetic studies have identified a particular genetic variant (referred to as “rs8034191”) that increases the risks of lung cancer, nicotine dependence, and associated smoking behaviour. However, there remains debate as to whether the association of this variant with lung cancer is direct or is mediated by pathways related to smoking behaviour.

The rs8034191 genetic variant is located adjacent to the *CHRNA5* gene region, which encodes a nicotine receptor. Nicotine is the major addictive component of cigarettes.

We drew 1,836 cases and 1,452 controls from a case-control study assessing the molecular epidemiology of lung cancer. Interviewer-administered questionnaires collected information on sociodemographic variables from each subject, including smoking intensity (number of cigarettes/day), and duration of smoking (years).

Regression models for lung cancer and for smoking intensity can be combined to calculate indirect effects mediated by smoking and direct effects through other pathways. The total effect is estimated by regression of lung cancer on the number of variant alleles for the genetic variant. The indirect effect is estimated by first regressing lung cancer on smoking intensity, and regressing smoking intensity on the genetic variant, then multiplying the effect estimates together. The direct effect is calculated as the total effect minus the indirect effect. All regression analyses are adjusted for age, sex, college education, and smoking duration.

Mediation analyses indicated strong evidence for a direct effect and suggested that the indirect effect is small. The direct-effect odds ratio was 1.35 (95% confidence interval (CI): 1.21, 1.52; $P = 3 \times 10^{-7}$), and the indirect-effect odds ratio was 1.01 (96% CI: 0.99, 1.02; $P = 0.15$) per additional copy of the variant allele for rs8034191. Analyses assume that conditional on the covariates, there is no confounding of 1) the exposure-outcome relation, 2) the mediator-outcome relation, or 3) the exposure-mediator relation and that 4) there is no effect of the exposure that itself confounds the mediator-outcome relation.

Interaction analyses were also performed to assess whether the association of the rs8034191 genetic variant with lung cancer risk varied in smokers and non-smokers. The test for interaction was significant on the additive risk scale ($P = 1 \times 10^{-3}$). The association between the variant and lung cancer risk was significant in smokers, but not in non-smokers.

Adapted and shortened from “Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interation” by VanderWeele et al, Am J Epidemiol 2012; 175(10):1013-1020.

3 Statistics in Medical Practice

A placebo-controlled clinical trial is conducted to estimate the effectiveness of a new drug designed to improve physical functioning and quality of life (QoL) in patients with a disabling disease. Five hundred patients are recruited, randomised with equal probability to receive the new drug or a placebo, and then followed up for 12 months. Each month the patients attend a specialist clinic, where various outcomes of interest are measured, including QoL. One hundred and seventy patients drop out during follow-up, i.e. they decide they will no longer attend subsequent clinic visits. The remaining 330 patients attend all 12 clinic visits. The trial investigators want to estimate the effect of the new treatment (compared to placebo) on various outcomes at 12 months after baseline, but here we shall focus on the outcome QoL.

1. State what it means for the data to be monotone missing. (Note that throughout this question the term 'the data' refers to the data on QoL and treatment assignment.)

2. Suppose that the data are monotone missing.

   (a) Defining all your notation carefully, write down a mathematical expression for the assumption that the data are 'missing at random' (MAR).

   (b) Explain in words what MAR means in this trial where the data are monotone missing.

   (c) Briefly discuss whether the assumption that the data are MAR is plausible for this clinical trial.

3. (a) Explain what it means for the missingness to be 'ignorable'.

   (b) Prove that if the data are MAR, then the missingness is ignorable.

4. As part of their standard care, all patients with this disabling disease in the population receive a monthly check-up visit from a nurse. At the end of each check-up visit, the nurse records in an electronic database his/her personal rough assessment of the patient's degree of disability. Suppose that the trial investigators were able to access the information on these rough assessments.

   (a) How could they use this information to improve the precision of their estimate of the effect of the new treatment (compared to baseline) on QoL at 12 months after baseline?

   (b) Apart from improving precision, what would be another advantage of using this information in this way?

4 Analysis of Survival Data

(a) Explain how to construct an *empirical likelihood* function for estimating the survivor function from a time-to-event dataset comprising individuals with observed events and individuals who are right-censored.

What is meant by *left-truncation*? Why is it important to take account of left-truncation when estimating a survivor function? What contribution does a left-truncated individual make to the empirical likelihood function?

(b) The following table shows the ages at which a set of Professors of Mathematics became a professor, together with their age at death. A '+' following the age of death indicates a right-censored observation.

Professor	Age at Election	Age at Death
Branestawm	57	62
McGonagall	35	67
Moriarty	18	74+
Plum	46	75
Proton	48	79+
Prune	53	81+

(i) Why is this dataset not suitable for estimating the survivor function for *time-from-birth-to-death* but is suitable for estimating the survivor function for *time-from-age-60-to-death*?

The remainder of this question refers only to the variable T, defined as time-from-age-60-to-death, and its survivor function $\mathrm{F}(t)$.

(ii) How many independent variables would you expect the empirical likelihood for $\mathrm{F}(t)$ to be a function of? Obtain the empirical likelihood function in terms of these variables.

(iii) Hence or otherwise estimate the median of T.

(iv) It is subsequently discovered that Professors Plum, Proton and Prune had lied about their birthdate, adding 20 to the year. All their ages in the table above need therefore to have 20 added to them. Obtain the empirical likelihood function corresponding to the corrected data.

(v) Hence or otherwise obtain an estimate of the median of T using the corrected data.

(vi) Explain why ascertaining the ages at death of Professors Proton and Prune does not affect the estimate of the median obtained from the empirical likelihood but do affect the estimate obtained from fitting an exponential distribution.

5 Analysis of Survival Data

(a) What is meant by the terms *hazard multiplier* and *baseline hazard* in the context of *proportional hazards* modelling? Describe how you would construct a likelihood function for the hazard multipliers in the case where the baseline hazard function is unspecified. Outline how you would obtain a non-parametric estimate of the integrated baseline hazard.

What is a *stratified* proportional hazards model? How does the division of a time-to-event dataset into strata affect the construction of the likelihood and the estimation of the integrated baseline hazard?

A *matched-pair* model is a stratified model with precisely two individuals in each stratum. Describe carefully how you would construct the likelihood for a matched-pair proportional hazards model. Can a reasonable estimate of the integrated baseline hazard be obtained?

(b) Patients with a congenital disease of both eyes were enrolled into a clinical trial of a new treatment. An eye was chosen at random and treated, the other eye acted as a control. The response variable was the time to worsening of symptoms for each eye.

Let the lth patient have recorded time $x_l^{(i)}$ for the control ($i = 0$) and the treated eye ($i = 1$) respectively, with indicator $v_l^{(i)}$ equalling 1 if $x_l^{(i)}$ is a time of symptom worsening and equalling 0 if $x_l^{(i)}$ is a censoring time.

There were twenty patients enrolled into the trial:

- of the 12 patients with $x_l^{(0)} < x_l^{(1)}$, there were eight with $v_l^{(0)} = 1$ and six with $v_l^{(1)} = 1$
- of the 8 patients with $x_l^{(0)} > x_l^{(1)}$, there were seven with $v_l^{(0)} = 1$ and four with $v_l^{(1)} = 1$

Let β be the hazard multiplier such that the hazard function for the control eye of the lth patient is $h_l(t)$ and for the corresponding treated eye is $\beta h_l(t)$. Find the maximum likelihood estimate of β.

What could be the difficulties if the eye to be treated is chosen by some means other than randomization, such as the treated eye is the left eye or the treated eye is the eye with worse vision?

6 Analysis of Survival Data

(a) What is a *frailty* random variable in the context of time-to-event analysis? Show how a *proportional frailty* model can be set up using a frailty variable U and a baseline hazard function $h_0(t)$. Why is it often possible to choose U to have a specific expectation? Why is it an advantage if $\mathbb{E}U$ can be chosen to equal 1?

Suppose $U \sim \texttt{exponential}(1)$ and $h_0(t) = \lambda$. Calculate the population survivor function and the population hazard. Comment on the form of the population hazard when (i) $t = 0$ and (ii) $t \to \infty$.

(b) Patients receiving the standard treatment for a certain cancer have an exponential time-to-death distribution with rate parameter β. If an experimental treatment is added to the standard treatment then there is a probability π that the patient's survival improves (the time-to-death distribution is now exponential with rate parameter γ, with $\gamma < \beta$) and a probability $1 - \pi$ that the patient's survival is unaffected (the time-to-death distribution remains exponential with rate parameter β). Whether or not a patient is subject to an improved survival distribution after receiving the experimental treatment can be considered an unobserved random variable.

Construct a proportional frailty model for the time-to-death distribution, defining the frailty random variable U such that $\mathbb{E}U = 1$. Obtain an expression for the ratio of the population hazard function for patients receiving the experimental treatment to that for patients receiving only the standard treatment. Comment on the form of the ratio when (i) $t = 0$ and (ii) $t \to \infty$. What are the implications for the design of a controlled trial investigating the experimental treatment?

END OF PAPER