UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday 12 June, 2017 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

A systematic review was carried out to compare the effectiveness of treatments for preventing stroke in patients who have previously experienced a severe stroke and are at high risk of recurrence. Data were extracted from 12 eligible trials: 6 trials compared warfarin and placebo; 4 trials compared aspirin and placebo; 2 trials compared warfarin and aspirin. The table shows the data on strokes occurring during the first 5 years after randomisation, extracted from three of these trials.

Trial name	Warfarin (Strokes / Total)	Placebo (Strokes / Total)	Log odds ratio	Variance of log odds ratio
Cohen 1989	10/60	70/120	-1.95	0.15
Trial name	Aspirin (Strokes / Total)	Placebo (Strokes / Total)	Log odds ratio	Variance of log odds ratio
Meier 1986	25/75	50/100	-0.69	0.10
Trial name	Warfarin (Strokes / Total)	Aspirin (Strokes / Total)	Log odds ratio	Variance of log odds ratio
Waterhouse 1992	40/300	110/300	-1.33	0.04

(a) Use indirect evidence to estimate a log odds ratio and its variance, comparing warfarin against aspirin. State in words the assumption made when performing this calculation. Calculate an approximate 95% confidence interval for the log odds ratio, based on indirect evidence. Interpret the log odds ratio estimate and its confidence interval in words. [You may assume that the 97.5% quantile of the standard Normal distribution is approximately equal to 2.]

(b) Give the formula for a pooled estimate and variance for the log odds ratio comparing warfarin against aspirin, based on both direct and indirect evidence.

(c) Show how to test for a difference between the direct and indirect evidence regarding the comparison of warfarin and aspirin.

(d) Give three reasons which could motivate researchers to carry out a network meta-analysis rather than presenting results from separate pairwise meta-analyses (in a generic setting rather than this specific example).

(e) The systematic review authors decide to perform separate pairwise meta-analyses for each treatment comparison. For the warfarin vs. placebo comparison, the trials were carried out in six different countries and definitions of eligibility for inclusion differed across trials. Write down a suitable model for combining the data from these six trials, and explain why this model is appropriate. Show how to assess the influence of the Cohen 1989 trial on the combined log odds ratio comparing warfarin and placebo.

2 Statistics in Medical Practice

(a) Multiple testing in GWAS

(i) In the context of testing the hypothesis that a biomarker is associated with bowel cancer, please define what a *Type I error* would be.

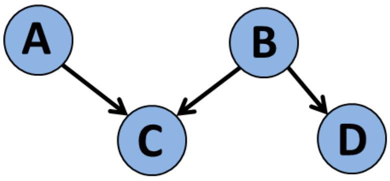
(ii) Suppose we test 10 biomarkers for association with bowel cancer, and declare significance according to a p-value threshold of 0.05. Assuming none of the biomarkers are associated with bowel cancer, and the biomarkers are not correlated with one another, what is the probability of obtaining one or more false positive results? You can leave your answer as a mathematical expression. Is this probability larger or smaller if the biomarkers are all positively correlated? Please provide a justification for your answer.

(iii) Suppose we have tested 100,000 genetic markers for association with body mass index (BMI). The p-values of association for the 10 markers most strongly associated with BMI are shown in the table below. Controlling for a family wise error rate of 0.05 and using a Bonferroni significance threshold adjusted for the 100,000 tests, how many genetic markers would be declared significant? How many genetic markers would be declared significant using the Benjamini and Hochberg method to control the false discovery rate at 10%?

Genetic marker	p-value
1	2.4E-08
2	7.2E-08
3	5.2E-07
4	8.2E-07
5	1.8E-06
6	3.8E-06
7	5.5E-06
8	6.0E-06
9	6.6E-06
10	1.1E-05

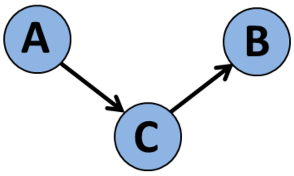
(b) Network structure learning

(i) A, B, C and D are random variables. Given the Bayesian network shown below, factorise the joint distribution $P(A, B, C, D)$ into a product of local probability distributions. Now assume A, B, C and D are binary. What is the minimum number of parameters required to define the factorised joint probability distribution?



(ii) For the Bayesian network shown below, decide which one of the following two statements is true and provide a proof [a counter example is not required for the false statement]:

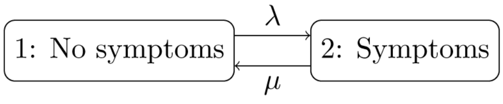
- $A \perp B$ (A is marginally independent of B)
- $A \perp B \,|\, C$ (A is conditionally independent of B given C)



(iii) Given data for p Gaussian random variables, $X_1,\ldots,X_p$, we would like to use the Bayesian score approach to find the Bayesian network structure that best represents the dependencies between the variables. Provide an overview of the procedure to calculate the Bayesian score (up to proportionality) for a given network structure G. Your answer should include an expression for the score (using Bayes' rule), an expression showing the marginalisation of nuisance parameters and an explanation of the benefit of using a conjugate parameter prior. Define any notation that you introduce and describe all terms in any expressions that you write down.

3 Statistics in Medical Practice

People with a chronic illness experience periods of symptoms, and we model this with a two-state continuous-time Markov model with symptom onset rate λ and recovery rate μ.



You may use without proof the following expressions for the transition probabilities over a time interval of t years:

$$\begin{aligned} p_{11}(t|\lambda,\mu) &= (\mu+\lambda E)/(\lambda+\mu) \\ p_{12}(t|\lambda,\mu) &= (\lambda-\lambda E)/(\lambda+\mu) \\ p_{21}(t|\lambda,\mu) &= (\mu-\mu E)/(\lambda+\mu) \\ p_{22}(t|\lambda,\mu) &= (\lambda+\mu E)/(\lambda+\mu) \end{aligned}$$

where $\mathrm{E} = \exp(-(\lambda+\mu)t)$.

(a) (i) What is the probability that somebody currently with no symptoms will spend the whole of the next t-year period without symptoms?

(ii) Obtain the expected total time spent with symptoms, over the t-year period, for someone currently with no symptoms, as a closed-form function of λ and μ.

(iii) Thus, in the limit as t increases, what is the proportion of time spent with symptoms?

(b) Suppose a set of people with the illness are examined for symptoms at intervals of exactly one year. Let n_{rs} be the number of times, in the whole dataset, when a person was observed in state r in one year followed by state s exactly one year later.

(i) Write down an expression for the joint log-likelihood for the transition intensities λ and μ given these data.

(ii) Suppose $\lambda = \mu$. Derive a condition in terms of n_{rs} for the likelihood to have a maximum over valid finite values of λ.

(iii) In practice, why might we not expect to have observations at equally-spaced intervals? Briefly discuss the consequences of this for maximum likelihood estimation.

(c) Suppose that some patients are given a treatment which aims to alleviate symptoms. The model was fitted to data, now allowing $\lambda \neq \mu$, and including a covariate representing this treatment. The maximum likelihood estimates are as follows.

	Transition intensities for untreated patients (per year)	Hazard ratios for treatment
No symptoms → symptoms	$\hat{\lambda} = 0.2$	$\exp(\hat{\beta}_\lambda) = 0.8$
Symptoms → no symptoms	$\hat{\mu} = 0.25$	$\exp(\hat{\beta}_\mu) = 2$

(i) Under this model, what is the expected length of a period of symptoms for a treated patient?

(ii) Use the answer from a)iii) to calculate the relative proportion of time spent with symptoms (in the long-term limit) for a treated person compared to an untreated person. Express the answer as a simple fraction.

(iii) Explain precisely how to perform a single test for the null hypothesis that the treatment has no effect on the course of the illness.

4 Analysis of Survival Data

(a) What is meant by the *survivor function* of a continuous time-to-event variable T? How would you obtain the *median* of T from its survivor function $\mathrm{F}(t)$?

A continuous time-to-event random variable T has survivor function $\mathrm{F}(t)$ such that $\mathrm{F}(t) = 0$ for $t > t^*$. Obtain an expression for the mean of T. Suggest a condition on $\mathrm{F}(t)$ sufficient for T to have a finite mean if, instead, $\mathrm{F}(t) > 0$ for all $t > 0$.

(b) Outline how to construct the *Kaplan-Meier* estimator of the survivor function $\mathrm{F}(t)$ 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.

How would you obtain an estimate of the median of a survival distribution from a Kaplan-Meier curve?

For a particular dataset, the Kaplan-Meier estimate of $\mathrm{F}(t)$ at $t = t_0$ has value ϕ_0. Immediately after t_0 there is just one individual in the risk set. What is the Kaplan-Meier estimate ϕ_1 of $\mathrm{F}(t_1)$, $t_1 > t_0$ in the cases:

(i) the remaining individual is censored at t_1;

(ii) the remaining individual has an event at t_1.

Comment on how reliable these estimates would be in practice. Can the Kaplan-Meier curves obtained in the two cases be used to estimate the mean of the time-to-event variable?

5 Analysis of Survival Data

Give a brief description of *proportional hazards* modelling, including definitions of the terms *hazard multiplier* and *baseline hazard.* Derive the partial likelihood function of a proportional hazards model in terms of the hazard multipliers [you may assume no ties in the dataset].

A proportional hazards model is being fitted to a certain dataset. The four individuals $\{a, b, c, d\}$ in the risk set at time $t = t_a$ have hazard multipliers ϕ_a, ϕ_b, ϕ_c, ϕ_d respectively. Individual b is right-censored at time t_b and the other three individuals have events at t_a, t_c, t_d respectively, with $t_a < t_b < t_c < t_d$. All other individuals in the dataset have events or are right censored at times strictly less than t_a.

(a) Write down the contribution to the partial likelihood from those four individuals.

(b) Suppose the event time for individual b would have been t^*. Write down the three possible orderings of t^*, t_a, t_c, t_d consistent with individual b being censored at t_b. Show that the sum of the three partial likelihoods obtained from the three orderings equals the partial likelihood obtained in part (a) and comment on this result.

Outline how the *Nelson-Aalen* estimator can be modified to provide an estimate of the integrated baseline hazard function.

(c) Using the same dataset as in parts (a) and (b), obtain an estimate for the increment $\mathrm{H}_0(t_d) - \mathrm{H}_0(t_a)$ in integrated baseline hazard in terms of the maximum partial likelihood estimates of the hazard multipliers.

6 Analysis of Survival Data

What is meant by a *competing risks* model? Define *cause-specific hazard* and *cumulative incidence function.* Derive the cumulative incidence function for a particular event in terms of the cause specific hazards.

A population receiving treatment for a disease is at risk of two types of mutually exclusive event: event A is death from the disease and event B is death as a side effect of the treatment. The cause-specific hazard for event A is constant over time and has value θ_A. The cause-specific hazard for event B is constant for time $t \leqslant \tau$ with value θ_B and is zero for $t > \tau$.

(a) Obtain the cumulative incidence function for event A.

(b) Obtain the cumulative incidence function for event B.

(c) Interpret the sum of the two incidence functions. Calculate the limit of the sum as $t \to \infty$ and comment.

(d) Given that an event (of either kind) has occurred before time τ what is the probability that the event is of type A?

7 Analysis of Survival Data

What is meant by a *frailty* model? Write down expressions for the *individual* and *population* survivor functions ($\mathbb{P}[T > t|U = u]$ and $\mathbb{P}[T > t]$ respectively) when the *individual* hazard function is given by

$$\mathrm{h}(t|U = u) = uh_0(t)\,,$$

where U is non-negative with density function $g(u)$ and $h_0(t)$ is the baseline hazard.

(a) In the case of a population where $U \sim \texttt{exponential}(1)$ and $h_0(t) = \theta t$ calculate the population survivor function and the population hazard function. Contrast how the individual hazard function and the population hazard function vary with time, and comment.

(b) Obtain the density of U conditional on $T > t$. Calculate $\mathbb{E}[U|T > t]$ and interpret.

(c) A second population also has $U \sim \texttt{exponential}(1)$ but the baseline hazard is now given by $h_0(t) = \lambda\theta t$ with $\lambda > 1$. Calculate the ratio of the population hazard functions for the two populations. Interpret the value of this ratio for $t = 0$ and as $t \to \infty$. Comment on the implications for proportional hazards modelling.

END OF PAPER