UNIVERSITY OF CAMBRIDGE

MAMA/207, NST3AS/207, MAAS/207

MAT3

MATHEMATICAL TRIPOS Part III

Tuesday 4 June 2024 1:30 pm to 4:30 pm

PAPER 207

STATISTICS IN MEDICINE

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

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

Over the past n days, there has been a rapid outbreak of a novel respiratory disease. To understand its potential to spread through the population, estimates of the time-varying reproduction number are required. Suppose that we can directly observe new infections (or a fixed fraction of them) as they occur, with daily rate Δ_t on day t. A time-since-infection model is implemented to link Δ_t, $t = 1, \ldots, T$ to the instantaneous infection rate $\beta_{t,\tau}$.

(a) Define the subscripts of the term $\beta_{t,\tau}$, and write down the discrete-time recursion linking Δ_t and $\beta_{t,\tau}$.

(b) There are multiple possible definitions of the reproduction number.

(i) Define both the *instantaneous* and *effective* reproduction numbers in terms of the infection rate. Identify which of these reproduction numbers is easier to estimate in practice.

(ii) Assume *separability* of the infection rate $\beta_{t,\tau}$. Obtain the relationship between the reproduction number chosen in (i) and the discretised generation interval distribution $\mathbf{g} = (g_1, g_2, \ldots,)$.

(iii) As we are in the early stages of an outbreak, it may be reasonable to assume that the underlying incidence is growing exponentially with rate ρ. Use your definition of the reproduction number to establish a relationship between the time-t reproduction number and the growth rate. If $g_i = \mathbb{P}(X = i-1)$, where $X \sim \text{Geometric}(p)$, show that the reproduction number can be expressed as:

$$\frac{1 - (1-p)\, e^{-\rho}}{pe^{-\rho}\left(1 - (1-p)^t\, e^{-\rho t}\right)}$$

In practice, the Δ_t are unobserved. However, noisy daily counts of the number of people diagnosed with the novel infection, $I_1, \ldots, I_T$ are available.

(c) Conditional upon the reproduction numbers and the generation interval, these count data can be assumed to be independent and Poisson-distributed.

(i) Write down a probability model that will give the joint distribution of the I_t, $t = 1, \ldots, T$.

(ii) Assume, *a priori* that the time-t reproduction number has a $\Gamma(\alpha_0, \beta_0)$ prior. Show that it has posterior mean

$$\frac{I_k + \alpha_0}{s_k + \beta_0}$$

for a suitably defined quantity s_k.

[QUESTION CONTINUES ON THE NEXT PAGE]

A cohort of people who have tested positive for the infection are followed up through time. We want to estimate the risk of death, and the expected times to death or recovery (defined as the first time a person would test negative if tested).

(d) Describe a continuous-time multi-state model that could be used to estimate these quantities, indicating the states and permitted transitions in continuous time, and defining symbols to represent the model parameters.

(e) Under this model, what is

(i) the probability that an infected person will die,

(ii) the expected time to the first event (either death or recovery)?

People are tested for the infection weekly, and for those who die we record the exact time of death. We observe the following data for two people.

Person 1	0 days	Test positive
	7 days	Test positive
	14 days	Test negative
Person 2	0 days	Test positive
	7 days	Test positive
	10 days	Death

(f) (i) Define the *transition probability matrix* of a continuous-time multi-state model.

(ii) Define the contribution to the likelihood of the multi-state model given by these two people, in terms of the transition probabilities.

(g) Explain an assumption being made in the model used in (d)–(f), and explain why it is challenging to fit such a model to this sort of data under a less restrictive assumption.

Suppose we extended the multi-state model to include a state which indicates that an infected person is in hospital. We have a population of N people, and the chance that a person gets infected (over some period of time) is p.

(h) Explain how we could estimate the total expected number of days spent in hospital resulting from infections among this population during this period. (Assume that a person cannot be infected repeatedly.)

2 Statistics in Medical Practice

Consider a clinical trial where a new treatment will be tested against a standard treatment. When treatment k is given to a patient (where $k = 0$ denotes the standard treatment and $k = 1$ the new treatment), a Bernoulli(p_k) outcome is observed, where p_k represents the probability of a treatment success. Denote the total number of patients on treatment k as n_k. The parameter of interest is the difference in the treatments' success rates $\delta = p_1 - p_0$ and the null hypothesis $H_0 : \delta = 0$ is to be tested.

(a) Define the Wald test statistic and write down its approximate (large-sample) normal distribution i) under H_0 and ii) when $\delta = \delta^*$ (where $\delta^* > 0$).

(b) Assume that equal numbers of patients are allocated to each treatment arm (so that $n_0 = n_1 = n$) and that n is large so that the approximate normal distribution derived in (a) holds. Derive a formula for the sample size required to detect a clinically relevant difference δ^* between the standard treatment and the new treatment, with power $1 - \beta$ and a type I error rate of α when using a one-sided Wald test.

(c) The clinical investigator for the trial has heard about group sequential designs and asks if such a design can reduce the sample size of the trial. What answer would you give?

(d) Consider redesigning the original trial using a two-stage group sequential design. Let Z_j denote the Wald test statistic at the jth analysis and m_j denote the cumulative group-size per arm. Write down the asymptotic joint distribution of Z_j in terms of δ, p_j and m_j.

(e) Using the joint distribution of (Z_1, Z_2), and given lack-of-benefit boundaries (l_1, l_2) and efficacy boundaries (u_1, u_2), derive an expression for the asymptotic probability of making a type I error.

(f) The clinical investigator has also heard about response-adaptive randomisation (RAR) and wonders about using RAR instead of a group sequential design. Give one potential advantage and one potential disadvantage of using RAR in a clinical trial.

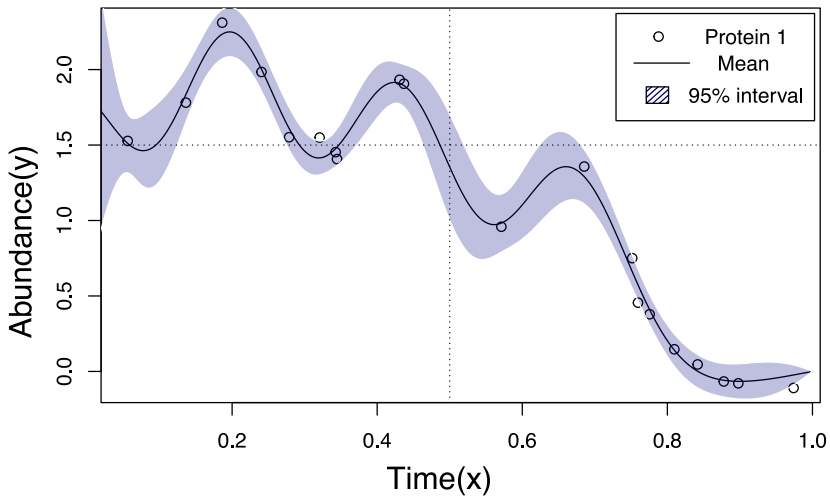
(g) The clinical investigator mentions historical data that suggests $p_0 = 0.1$ and pilot data suggesting $p_1 = 0.5$. Assuming that these are the true values, calculate the large-sample variance of the estimated treatment difference $\hat{p}_1 - \hat{p}_0$ (as in (a)) when using a RAR design that targets the Neyman ratio (for a fixed total sample size $n_{\max} = n_0 + n_1$), and compare this with the variance achieved when using equal allocation.

3 Statistics in Medical Practice

In the figure below, the abundance of a protein in a cell has been plotted as a function of time over the duration of the cell cycle. Given $n = 20$ observations of the abundance at various time points during the cycle, a standard Gaussian process (GP) regression model was fit to the data according to the model:

$$y_j|f, x_j, \sigma^2 = f(x_j) + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad j = 1, \ldots, n \tag{1}$$
$$f \sim \mathcal{GP}(0, \kappa(\cdot, \cdot)).,$$

The parameters of the model (kernel hyperparameters and the noise variance) can be considered fixed throughout the question. In the figure, the original data are plotted alongside the posterior mean of the GP and a 95% credible band.



(a) Given the observed data $\mathbf{y}$ at locations $X = \{x_1, \ldots, x_n\}$, derive the predictive posterior distribution of the GP at an unobserved input x_*, i.e. $f(x_*)|\mathbf{y}, X, x_*$.

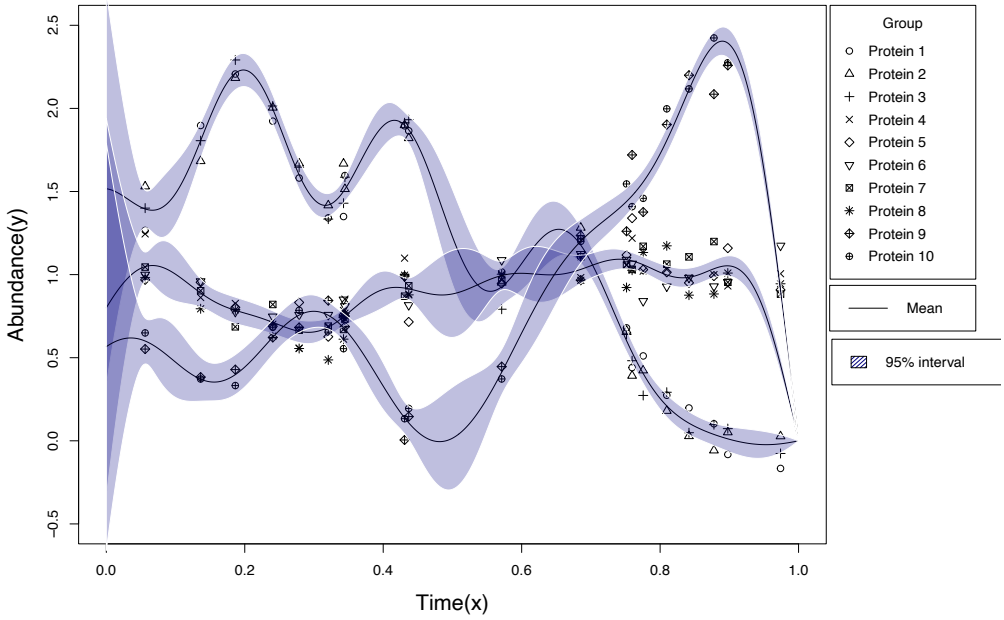
Make sure that the dimensions and entries of any vectors or matrices you define are explicitly specified. [*Note that the conditioning property of the multivariate normal is given at the end of the question.*]

(b) For this particular experiment, the researchers are interested in the abundance of the protein at the midway point of the cell cycle, i.e. $f(0.5)$.

Use the solution arrived at in (a) to find an explicit expression for the posterior probability that the abundance is below 1.5, i.e. $P(f(0.5) < 1.5|\ \mathbf{y}, X)$.

[QUESTION CONTINUES ON THE NEXT PAGE]

Consider now an extension of the previous experiment, where we are given the abundances of 10 different proteins over the course of the cell cycle. The figure below illustrates that the behaviours of these proteins can be well described by only 3 functions, which we denote by f_1, f_2, and f_3.



Each protein is observed at a common, finite set of time points, $X = \{x_1, \ldots, x_n\}$, and we denote by $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})^T$ the corresponding vector of measured abundances for the i-th protein. For $m = 1, 2, 3$, we denote by $\mathbf{f}_m = (f_m(x_1) \ldots, f_m(x_n))^T$ the vector whose entries are given by the m-th function evaluated at each time point. For the i-th protein, we assume that the measured abundance is generated (noisily) from one of the 3 functions, i.e. $y_{ij} = f_m(x_j) + \epsilon_{ij}$ for some m. Furthermore, we assume that the functions f_1, f_2, f_3 were drawn *iid* from a zero-mean GP with kernel $\kappa(\cdot, \cdot)$.

We can define a mixture model for our data as follows:

$$\begin{aligned} \pi_1, \pi_2, \pi_3 | \boldsymbol{\alpha} &\sim Dir(3, \boldsymbol{\alpha}) \\ \mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3 | x_1, \ldots, x_n &\overset{iid}{\sim} \mathcal{N}(\mathbf{0}, K) \\ \mathbf{y}_1, \ldots, \mathbf{y}_{10} | \{\pi_k, \mathbf{f}_k\}_{k=1}^3, \sigma^2 &\overset{iid}{\sim} \sum_{m=1}^{3} \pi_m \phi(\mathbf{y}; \mathbf{f}_m, \sigma^2), \end{aligned} \tag{2}$$

where $\phi(\mathbf{y}; \mathbf{f}_m, \sigma^2) = \mathcal{N}(\mathbf{f}_m, \sigma^2 I)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$, and K has entries $K_{ij} = \kappa(x_i, x_j)$.

[QUESTION CONTINUES ON THE NEXT PAGE]

(c) Let's assume for the time being that for each observation $\mathbf{y}_i$ we are given a label $c_i \in \{1, 2, 3\}$ indicating the function from which it was generated; i.e. $\mathbf{y}_i | c_i = m, \{\pi_m, \mathbf{f}_m\}_{m=1}^3, \sigma^2 \sim \phi(\mathbf{y}_i; \mathbf{f}_m, \sigma^2)$, where $P(c_i = m) = \pi_m$. We have two sets of unknown parameters in the model: the mean vectors $\{\mathbf{f}_m\}_{m=1}^3$ and mixture weights $\{\pi_m\}_{m=1}^3$.

To perform inference in this finite mixture model, we can employ a Gibbs sampler, where at each iteration, each parameter is sampled from its conditional posterior distribution given all the other parameters.

(i) Assuming that the vector $\boldsymbol{N} = (N_1, N_2, N_3)^T$ follows a multinomial distribution $M_3(10; \pi_1, \pi_2, \pi_3)$, where $N_m = |\{c_i : c_i = m\}|$, find the conditional posterior distribution for the mixture weights,

$$p(\pi_1, \pi_2, \pi_3 | \{\mathbf{y}_i, c_i\}_{i=1}^{10}, \{\mathbf{f}_m\}_{m=1}^3)$$

(ii) Find the conditional posterior distribution for the vector $\mathbf{f}_m$,

$$p(\mathbf{f}_m | \{\mathbf{y}_i, c_i\}_{i=1}^{10}, \{\pi_m\}_{m=1}^3) \text{ for } m = 1, 2, 3.$$

Make sure that the dimensions and entries of any vectors or matrices you define are explicitly specified. [For notational purposes, denote by $\mathbf{y}_1^{(m)}, \ldots, \mathbf{y}_{N_m}^{(m)}$, the observations generated by the m-th function (i.e. those where $c_i = m$).]

Suppose now that we are *not* given the labels, $\{c_i\}_{i=1}^{10}$.

(iii) Find the conditional posterior distribution for c_i,

$$P(c_i = m | \{\mathbf{y}_i\}_{i=1}^{10}, \{\pi_m\}_{m=1}^3, \{\mathbf{f}_m\}_{m=1}^3).$$

Finally, suppose we did *not* know the number of functions in advance.

(iv) Write up a hierarchical model akin to the model in (2) that also takes the unknown number of functions into account.

Multivariate normal conditioning property Let $\mathbf{x} = \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix}$ be a Gaussian random variable with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)^T$ and covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}$. Then the conditional distribution of $\mathbf{x}_A$ given $\mathbf{x}_B$, is also a Gaussian:

$$\mathbf{x}_A | \mathbf{x}_B \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{A}} + \Sigma_{AB}\Sigma_{BB}^{-1}(\mathbf{x}_B - \boldsymbol{\mu}_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}).$$

4 Analysis of Survival Data

(a) A time-to-event dataset comprises n individuals, the ith either having an event ($v_i = 1$) or being censored ($v_i = 0$) at time x_i, $i = 1, \ldots, n$. There are no ties in the dataset (that is: for $i \neq i'$ and $v_i = v_{i'} = 1$, $x_i \neq x_{i'}$).

Each individual belongs to one of two groups, with the group membership of the ith individual indicated by g_i with $g_i \in \{0, 1\}$. The time-to-event distribution is the same for all individuals in a group. It is required to test whether the time-to-event distribution for group 0 is the same as that for group 1.

Let a_j, $j = 1, \ldots, m$ be the set of times at which there is an event, with $0 < a_{j-1} < a_j$.

(i) What is meant by a *risk set*? Write down an expression for the number of individuals $r_j^{(k)}$ in the risk set for group k at time a_j.

(ii) Let the random variable U_j, $j = 1, \ldots, m$ indicate the group membership of the individual who has an event at time a_j. Derive the expectation $\mathbb{E}U_j$, conditional on the observed $r_j^{(0)}$ and $r_j^{(1)}$, assuming that the time-to-event distributions of the two groups are equal.

(iii) Interpet the expression $u_j - \mathbb{E}U_j$ where u_j is the observed group membership of the individual having the event at time a_j.

(iv) Hence construct a statistic T_0 which can be used to test the null hypothesis that the two groups have the same time-to-event distribution. [You need not normalize to unit variance.]

(b) Give a brief derivation of the *Nelson-Aalen* estimator of the integrated hazard function.

(c) This part of the question refers to the dataset defined in part (a).

(i) What is the Nelson-Aalen estimate $\hat{\mathrm{H}}_j^{(k)}$ of the integrated hazard at time a_j for group k?

(ii) Write down an expression for the increment $\Delta\hat{\mathrm{H}}_j^{(k)}$ at time a_j in the estimate of group k's integrated hazard, that is: $\Delta\hat{\mathrm{H}}_j^{(k)} = \hat{\mathrm{H}}_j^{(k)} - \hat{\mathrm{H}}_{j-1}^{(k)}$, with $\hat{\mathrm{H}}_0^{(k)}$ defined equal to 0.

(iii) A general procedure to obtain a statistic T_W to test the equality of two integrated hazard functions is to construct a weighted sum over the a_j of the difference in the increments in the estimated integrated hazards between the two groups. that is:

$$T_W = \sum_{j=1}^{m} w_j \left(\Delta\hat{\mathrm{H}}_j^{(1)} - \Delta\hat{\mathrm{H}}_j^{(0)} \right).$$

Derive the form of w_j which makes the two test statistics T_W and T_0 (derived in part (a)) equivalent.

(iv) Comment on why the weights you have calculated result in a better test statistic than simply setting $w_j = 1$ for all j.

5 Analysis of Survival Data

(a) What is meant by a *proportional hazards* family of time-to-event distributions?

(i) Show that if two continuous time-to-event distributions have survivor functions F_1, F_2 related by $\mathrm{F}_2(t) = (\mathrm{F}_1(t))^\lambda$ for some $\lambda > 0$ then they belong to the same proportional hazards family.

(ii) Let two time-to-event variables T_1, T_2 be of form:

$$\log T_k = \log a_k + b \log U_k$$

for $k \in \{1, 2\}$ with $a_k > 0$, $b > 0$ and U_k a random variable with `exponential(1)` distribution. Show that T_1, T_2 belong to the same proportional hazards family.

For the rest of this question, you may assume a dataset of form $\{(x_i, v_i, z_i)\colon i = 1, \ldots, n\}$ comprising 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 and z_i indicating which of two groups that individual belongs to. The labels i have been allocated to individuals such that the x_i are ordered with no ties: $x_{i'} < x_i$ for $i' < i$.

(b) Describe briefly how you would use a fully parametric proportional hazards model to test the hypothesis that the two groups have the same time-to-event distribution.

(c) What is a *semi-parametric* proportional hazards model? What is meant by a *partial* likelihood?

(i) Describe carefully how you would use a semi-parametric proportional hazards model to test the hypothesis that the two groups have the same time-to-event distribution. [Do not attempt to maximize the partial likelihood.]

(ii) Suppose that $n > 3$, and that x_{n-2},x_{n-1},x_n correspond respectively to a censored observation and two observed events. Let t_{n-2} denote the actual (unobserved) event time for the $(n-2)$th individual.

Obtain an expression for the partial likelihood for each ordering of

$$x_1, \ldots, x_{n-3}, x_{n-1}, x_n, t_{n-2} \ ,$$

that could result from the different positions of t_{n-2} in the sequence

$$x_1, \ldots, x_{n-3}, x_{n-1}, x_n$$

which are consistent with the $(n-2)$th individual being censored at x_{n-2}.

Show algebraically that the sum of these partial likelihoods is equal to the partial likelihood calculated with the $(n-2)$th individual censored at x_{n-2}.

6 Analysis of Survival Data

(a) What is meant by *empirical likelihood*? Explain how to construct a non-parametric estimate of the survivor function by combining individual contributions to the empirical likelihood function. What properties of the survivor function can be used to simplify the maximization of the empirical likelihood function?

Write down the contributions to the empirical likelihood function from the individuals described below, explaining carefully your notation:

(i) A patient, taking part in a study of the time interval from randomization into a study of a new cancer treatment to death, who is still alive when the study is analysed 36 months after their randomization.

(ii) A patient, taking part in a study of the time interval from randomization into a study of a new cancer treatment to death, who dies on-study 42 months after their randomization.

(iii) A chimpanzee, being observed in a study of the time interval from sunrise to first awakening, who is seen to be already awake when the naturalist arrives at the study site 15 minutes after sunrise.

(iv) A patient, taking part in a study of the time interval from the diagnosis of a primary cancer to the first appearance of a secondary cancer, who is observed to have no secondary cancers 6 months after the diagnosis of the primary cancer but to have a secondary cancer 9 months after the diagnosis of the primary cancer.

(v) A bus, being observed in a study of the time interval from scheduled arrival to actual arrival, which is not seen to arrive by an observer who waits at the bus stop from 2 minutes after the scheduled time to 20 minutes after the scheduled time.

(vi) A person, taking part in a study of the time interval from 70th birthday to death – the study population being the inhabitants of a care home, who starts living at the care home on their 73rd birthday and dies on their 83rd birthday.

(b) What is meant by a *period* survival analysis? Describe how to construct a period survivor function for a specified time period, using – as an illustration – a period survival analysis for calendar year 2022 of time from diagnosis of a particular cancer to death from any cause.

Suppose you want to compare two such period survivor functions obtained for 2022 from two different countries. Explain briefly how you would make the comparison:

(i) non-parametrically;

(ii) parametrically, on the assumption that all patients within each country experienced an identical constant hazard throughout 2022.

END OF PAPER