UNIVERSITY OF CAMBRIDGE

MAMA/207, NST3AS/207, MAAS/207

MAT3

MATHEMATICAL TRIPOS Part III

Friday 5 June 2026 9:00am to 12:00pm

PAPER 207

STATISTICS IN MEDICAL PRACTICE

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 University Technical Colleges (UTCs) are specialist secondary schools in England focusing on technical and scientific education. There are currently 43 UTCs in England, and around 3500 secondary schools that are not UTCs (non-UTCs).

A government minister is considering whether to convert some existing secondary schools in England to become UTCs. She is interested in increasing the proportion of school leavers who take up apprenticeships.

(a) From a causal inference perspective, what quantity would you ideally want to estimate in order to understand if the minister's policy will be effective?

A policy adviser suggests comparing the average proportion of school leavers at UTCs and non-UTCs who take up apprenticeships.

(b) Is this a sensible suggestion? Justify your answer.

Using publicly available data, you perform a matched analysis comparing UTCs and non-UTCs. You match on various contextual data: percentage of students receiving free school meals (%FSM), percentage of students with English as an additional language (%EAL), percentage of students with special educational needs (%SEN), and percentage of boys (%BOYS). For each UTC, you select the five non-UTCs which are the closest match according to these covariates, and compare the average proportion of school leavers at UTCs and these matched non-UTCs.

(c) Is this a better analysis suggestion? Justify your answer.

A civil servant suggests that rather than performing a 1:5 matched analysis, that you perform a 1:50 matched analysis instead, matching each UTC with 50 non-UTCs.

(d) Provide and justify one advantage and one disadvantage of this suggestion.

(e) How could you assess whether your analysis is able to unbiasedly estimate the desired quantity? Give two suggestions, each with a justification.

Several UTCs have unbalanced sex ratios (that is, they have either disproportionately many boys, or disproportionately many girls).

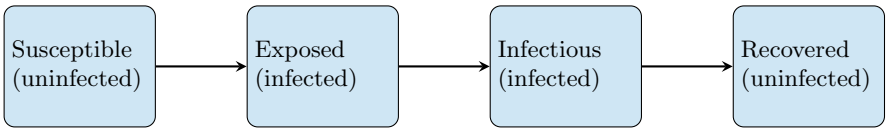
(f) How might this affect the validity of our matched analysis?

A statistical analysis can estimate an average treatment effect (ATE) parameter, representing the average effect of changing treatment status in all experimental units, or an average treatment effect in the treated (ATT) parameter, representing the average effect of changing treatment status in experimental units who receive the treatment.

(g) Which parameter would be more relevant to estimate in this investigation? Justify your answer.

2 Statistics in Medical Practice

We have a multistate model for an infectious disease, of the form:



We assume that a person cannot get infected twice, thus the Recovered state is absorbing.

In parts (a)–(c), we model the progression of individuals through these states with a continuous-time Markov multistate model with transition intensities q_{SE}, q_{EI}, q_{IR}.

(a) Suppose we follow up 1000 initially-uninfected people for 3 months each, and observe 10 infections. Obtain an estimate of q_{SE} by simple arithmetic, and state what you are assuming about how infections were observed. Suggest why this assumption might not be true in practice.

(b) Suppose we have a test which detects the infection, but it cannot distinguish between the Exposed and Infectious states. We test individuals at specific times. Given data of the following form from one individual: *"negative at 1 month; positive at 2 months; negative at 3 months"*, obtain their contribution to the likelihood for the transition intensities. (You may use expressions indicating transition probabilities without further evaluation.)

(c) Suppose we then obtained estimates and a covariance matrix for the (log) transition intensities. Explain how we would then derive an estimate and confidence interval for the time that a person without the infection (but susceptible) is expected to spend infectious over the next 3 months. (No analytic or numerical calculations need to be done to answer the question, but define as many quantities as you are able to, and indicate where numerical computation may be required in practice.)

Epidemic surveillance data are rarely at the level of the individual. In parts (d)–(f), we consider population-level models.

(d) Suppose that our 1000 individuals are part of a much larger susceptible population of size S_0 out of a total population of size N in which we can assume deterministic transmission dynamics. Assume that, initially, there is a batch of infectious individuals and that the initial number of exposed individuals is zero.

(i) In terms of $S(t), E(t), I(t), R(t)$, the number of people in the susceptible, exposed, infectious and recovered states, and using the parameters q_{SE}, q_{EI}, q_{IR} write down an ODE for the evolution of the system.

(ii) Write down the time-t expression for $S(t)$ that solves this ODE.

(iii) Multiply your equation for $\dot{E}(t)$ by an integration factor of $e^{q_{EI}t}$ to show that, for $q_{SE} \neq q_{EI}$:

$$E\left(t\right) = \frac{S_0 q_{SE}}{q_{EI} - q_{SE}} \left(e^{-q_{SE}t} - e^{-q_{EI}t}\right)$$

(iv) Find an expression for the time at which the number of exposed individuals is at a maximum.

(e) For an infectious disease such as influenza, the population-level model of (d) would be inappropriate.

(i) Identify which of the transitions $S \to E$, $E \to I$ and $I \to R$ would need to be redefined, and why.

(ii) Using a suitably defined parameter β, write out the more realistic ODE.

(f) In this new expression of the $SEIR$ transmission model, find an expression that links $S(t)$ and $R(t)$, and use this to show the epidemic 'final size' relation below, where S_∞ is the number of people who remain susceptible at the end of the epidemic:

$$\Rightarrow S_\infty = S_0 \exp\left(-\mathcal{R}\left(N - S_\infty\right)\right)$$

taking care to provide a definition of the quantity $\mathcal{R}$.

3 Statistics in Medical Practice

A Phase IIa trial investigates a novel therapy's binary clinical response at 6 months. The historical positive response rate is $\pi_0 = 0.20$. To protect patients from an ineffective therapy, a two-stage design is implemented:

- **Stage 1:** Enroll n_1 patients. If the number of positive responses $X_1 \leqslant r_1$, stop the trial and declare the drug futile.
- **Stage 2:** If $X_1 > r_1$, enroll an additional n_2 patients. At the end of the trial, reject $H_0 : \pi \leqslant 0.20$ if the total number of positive responses $X = X_1 + X_2 \geqslant r$.

(a) Let $b(k; n, p)$ denote the probability mass function and $B(k; n, p)$ denote the cumulative distribution function of a binomial random variable. Write down the exact mathematical expression for the **Overall Type I Error rate (α)** of this specific two-stage design.

A subsequent Phase IIb study compares two active doses (High Dose vs. Low Dose) with a binary outcome. Let p_H and p_L denote the true probabilities of response. A response-adaptive randomization (RAR) scheme is implemented.

(b) Write down the theoretical Neyman allocation ratio in terms of p_H and p_L. Then, (i) state the "ethically optimal" allocation ratio designed to minimize the expected number of treatment failures for a fixed variance for the Wald test for $p_H - p_L$, and (ii) discuss the difference between these two ratios.

(c) Briefly describe how adaptive allocation targeting either of these optimal ratios impacts the standard error of the estimated treatment difference $\hat{p}_H - \hat{p}_L$ compared to a fixed 1:1 scheme. State one major logistical or statistical challenge of using RAR in practice.

A Phase III trial randomizes patients 1:1 to either the gene therapy or placebo (n_T and n_C patients respectively) for exactly 1 year. The primary endpoint is the number of severe respiratory exacerbations. Exacerbation counts in the treatment (X_T) and control (X_C) groups are modeled as independent Poisson random variables with expected rates λ_T and λ_C.

The clinical team tests the hypothesis based on the **Rate Ratio (RR)**, testing $H_0 : \mathrm{RR} = 1$ versus $H_1 : \mathrm{RR} \neq 1$, where $\mathrm{RR} = \lambda_T/\lambda_C$.

[QUESTION CONTINUES ON THE NEXT PAGE]

(d) Using the maximum likelihood estimates $\hat{\lambda}_T = X_T/n_T$ and $\hat{\lambda}_C = X_C/n_C$, and applying the Delta method with a normal approximation, **show that** the Wald-type test statistic for the *log-transformed* rate ratio, ln(RR), is given by:

$$Z = \frac{\ln(\hat{\lambda}_T) - \ln(\hat{\lambda}_C)}{\sqrt{\frac{1}{n_T\hat{\lambda}_T} + \frac{1}{n_C\hat{\lambda}_C}}}$$

Explicitly state the asymptotic variance of ln(RR) under the null hypothesis ($\lambda_T = \lambda_C = \lambda$).

(e) Using the test statistic provided in (d), construct a formula to calculate the total required sample size ($N = n_T + n_C$) to detect a true rate ratio of RR_A with $1 - \beta$ power at a two-sided significance level of α. Assume equal allocation ($n_T = n_C = N/2$).

4 Analysis of Survival Data

A statistician requires a simulated sample of n observations from an $\texttt{Exponential}(\lambda)$ time-to-event distribution, where λ is the rate parameter. The simulated dataset is to be of form $\{(t_i, c_i, x_i, v_i)\colon i = 1, \ldots, n\}$ where t_i is the time-to-event, c_i is the time-to-censoring, x_i is $\min(t_i, c_i)$ and v_i is $\mathbb{1}[t_i \leqslant c_i]$.

(a) Given a source of random numbers from a $\texttt{Uniform}(0, 1)$ distribution, what transformation needs to be applied to give an $\texttt{Exponential}(\lambda)$ time-to-event variable?

(b) Suppose, initially, the simulated dataset is to contain no censored observations (that is: $c_i = \infty$). Write down the ratio of the expected total number of events $\sum_i v_i$ to the expected total time at risk $\sum_i x_i$.

(c) Suppose now that a sample of n observations $\{(t_i\colon i = 1, \ldots, n\}$ have been generated from a $\texttt{Exponential}(\lambda)$ distribution. The investigator wants to use this sample to construct a simulation of an $\texttt{Exponential}(\lambda)$ time-to-event distribution with a proportion q of the observations censored. Consider the three following possible censoring mechanisms:

(i) the censoring time c_i is independent of i: $c_i = \tau$.

(ii) the censoring time c_i is a sample from an $\texttt{Exponential}(\mu)$ distribution.

(iii) a sample u_i is drawn from a $\texttt{Uniform}(0, 1)$ distribution. t_i is declared to be a censoring time (that is: not an event time) if $u_i \leqslant \pi$.

For all three censoring mechanisms, state whether the censoring is informative or uninformative, justifying your answer.

Derive expressions for τ, μ and π which result in the expected proportion of censored events equalling q.

For all three mechanisms, calculate the ratio of the expected total number of events $\sum_i v_i$ to the expected total time at risk $\sum_i x_i$, and compare that ratio to that obtained in the case of no censoring. For any mechanism where the ratio is not the same as that for no censoring, suggest how to modify the simulation of t_i to achieve equality. Interpret your answer in terms of competing risk methodology.

5 Analysis of Survival Data

(a) Describe the Cox regression formulation of the proportional hazards model and explain how it can be used to test the hypothesis that two time-to-event distributions have the same hazard function (you may assume no tied event times in the data).

(b) Patients are enrolled in pairs into a clinical trial of a new cancer treatment, matched by age and general health, with one member of the pair randomised to active treatment ($k = 1$) and the other randomised to placebo ($k = 0$). The outcome variable is time to return of symptoms.

The time to return of symptoms or censoring for thirteen pairs is shown below in the format $(x_i^{(0)}, x_i^{(1)})$ with $x_i^{(k)}$ being the time to event or censoring for the patient in the ith pair receiving treatment k, a plus sign (+) indicating a censored observation:

(2, 106)	(5, 4)	(9, 59)	(11, 391+)	(17, 17)
(20, 95)	(21, 38+)	(46, 312)	(62, 99+)	(63+, 278)
(89+, 1083+)	(140, 140+)	(141, 138+)		

(i) Describe how to use a stratified Cox regression model to test the hypothesis that the active treatment has no effect on time to return to symptoms, relative to the placebo treatment.

Let $\exp(\beta k)$ be the hazard multiplier corresponding to treatment k:

(ii) which of the pairs of patients in the above dataset are *not* informative with respect to inference about β?

(iii) obtain an estimate for $\exp(\beta)$, and test the hypothesis that $\beta = 0$.

(c) What might be an advantage of using a matched-pair analysis over a comparison between two treatments ignoring the matching? What might be a disadvantage?

6 Analysis of Survival Data

(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. Assuming that all individuals experience the same constant hazard $\mathrm{h}(t) = \theta$, derive the maximum likelihood estimator $\hat{\theta}$ of θ.

(b) (i) What is meant by left truncation?

(ii) Describe how you would modify the Kaplan-Meier estimation of a survivor function in the presence of left-truncated individuals.

(iii) Describe how you would modify the estimator for $\hat{\theta}$ in part (a) in the presence of left-truncated individuals.

(iv) Explain briefly how the concept of left truncation can be used to assess cancer survival in a particular *period* of calendar time.

(c) An investigator has the collected the following dataset of time in days from first administration of a new treatment to death:

3	4	4	7	12	13+	15+	17	20	25+
78	105	133+	166	181	236	253	258	278+	392

where a '+' indicates a censored observation.

(i) The investigator believes that the hazard is piecewise constant: being θ_1 for $t \leqslant 28$ and θ_2 for $t > 28$. Obtain estimates for θ_1 and θ_2.

[You may find these sums over the dataset's observation times useful: $\sum_{i:x_i \leqslant 28} x_i = 120$ *and* $\sum_{i:x_i > 28} x_i = 2080$.*]*

(ii) Describe how to construct a likelihood ratio test of the hypothesis $\theta_1 = \theta_2$. Write down a numerical expression for the likelihood ratio statistic (do not attempt to evaluate that expression).

END OF PAPER