UNIVERSITY OF CAMBRIDGE

MAMA/207, NST3AS/207, MAAS/207

MAT3

MATHEMATICAL TRIPOS Part III

Monday, 12 June, 2023 9:00 am to 12:00 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

A recent study used cross-sectional data from the United States (US) to compare COVID-19 rates in households where a named individual in the household had a birthday in the previous two weeks (referred to subsequently as a "birthday event") versus households where the individual did not have a birthday event. The aim of the investigation was to assess the potential causal effect of social gatherings (such as birthday parties) on COVID-19 infection rates using birthday event as an instrumental variable.

1. Write clearly the instrumental variable assumptions in terms of the variables used in this study.

2. Do you think the instrumental variable assumptions are plausible in this investigation? Provide two suggestions how the instrumental variable assumptions could be violated, and two suggestions (potentially linked) for how the validity of the instrument variable assumptions could be assessed.

Investigators presented their results in categories stratifying for county-level prevalence of COVID-19 in Figure 1, where "first" refers to counties in the decile with the lowest levels of COVID-19, and "tenth" refers to counties in the decile with the highest levels of COVID-19. We see that the change in COVID-19 infection rates (that is, the difference in rates between households with and without a birthday event) is close to zero in the first and second deciles, but differs significantly from zero in other deciles. (A county is a geographic subunit of the US. Most US counties have a population of a few thousand people.) As a supplementary analysis, investigators compared the change in COVID-19 infection rates between households with a birthday event versus without a birthday event in "red" counties versus "blue" counties. "Red" counties were defined as those with a majority vote for a Republican presidential candidate in the 2016 US presidential election. "Blue" counties were defined as those with a majority vote for a Democrat presidential candidate in the 2016 US presidential election. Generally speaking, Democrat voters are more likely to take a cautious approach to COVID-19 compared with Republican voters.

3. Why might differences in the change in COVID-19 infection rates associated with a birthday event between "red" and "blue" counties be informative about the causal effect of social gatherings on COVID-19 risk? What assumptions are made in this supplementary analysis?

4. Provide two suggestions how investigators could improve the reliability of this supplementary analysis.

[QUESTION CONTINUES ON THE NEXT PAGE]

5. Suggest one further similar supplementary analysis that could be attempted that divides the US population into subgroups, and compares differences in the association between the instrument and COVID-19 infection rates between the subgroups. Provide brief justification:

   i) why this subgroup comparison is worthwhile,

   ii) what you might expect to see if there truly was a causal effect of social gatherings on COVID-19 infection rates,

   iii) what assumptions the analysis makes, and

   iv) how you may assess those assumptions.

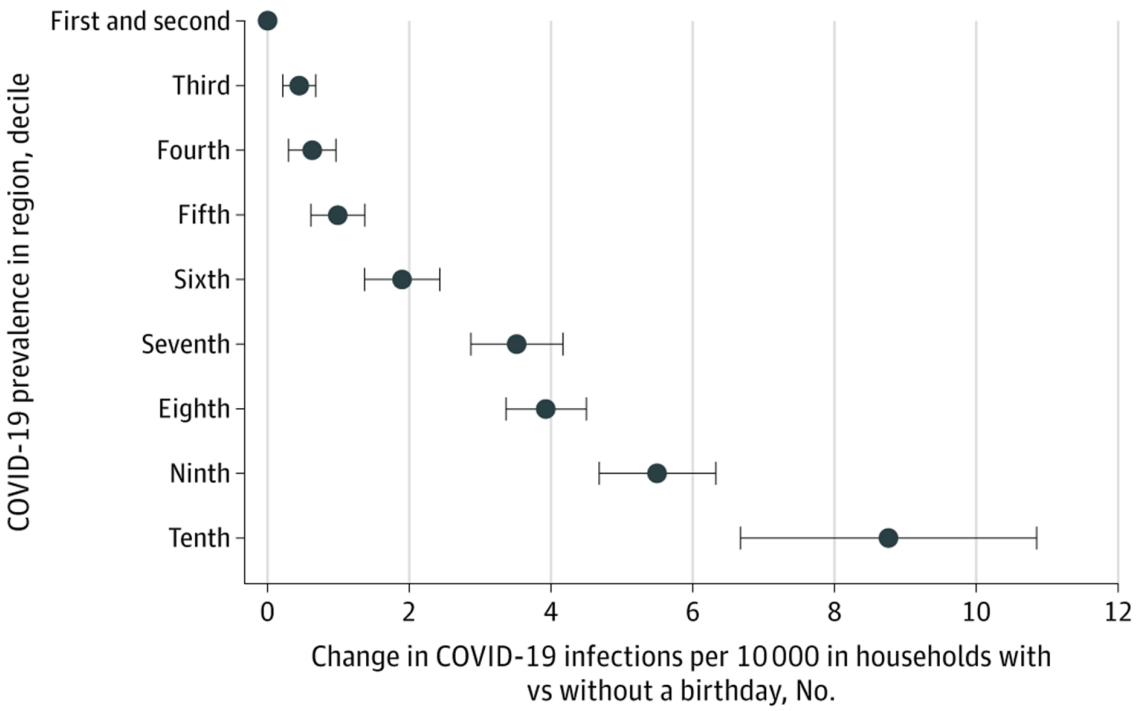

Figure 1: Absolute difference in COVID-19 rates per 10,000 individuals between households with and without a birthday event in deciles of the US defined by county-level baseline COVID-19 prevalence. Error bars represent 95% confidence intervals.
Taken from Whaley et al, JAMA Intern Med. 2021;181(8):1090-1099. doi:10.1001/jamainternmed.2021.2915.

2 Statistics in Medical Practice

Consider a two-armed clinical trial that tests a new treatment ($k = 1$) against a standard treatment ($k = 0$). A Bernoulli outcome is observed from each patient after receiving their treatment: $y_{ik} = 1$ indicates a response of patient i receiving treatment k, and $y_{ik} = 0$ indicates no response. Denote the response rate by p_k for treatment group $k = 0, 1$. The trialist is interested in performing hypothesis tests using the log relative risk of failure, denoted by $\theta = \log\left(\frac{1-p_1}{1-p_0}\right)$. Patients will be randomised to treatments using a response-adaptive randomisation procedure. Denote the total number of patients allocated to treatment k at the end of the trial by n_k (with $n_k \geqslant 1$) and the observed outcome from patient i on treatment k by y_{ik} for $i = 1, \ldots, n_k$. Let the allocation ratio be defined as $R = \frac{n_0}{n_1}$.

(a) The Z-test (Wald test) is based on a statistic $Z = \frac{\hat{\theta}}{\sqrt{\frac{p_0}{n_0(1-p_0)} + \frac{p_1}{n_1(1-p_1)}}}$. To compute $\hat{\theta}$, we use $\hat{p}_k = \frac{\sum_{i=1}^{n_k} y_{ik}}{n_k}$ for $k = 0, 1$.

(i) Formulate an optimisation problem to find the allocation ratio that maximises the power of the Z-test as a function of parameters p_0, p_1.

(ii) Show that given a fixed sample size $n = n_0 + n_1$, such an allocation ratio is equal to $R^* = \sqrt{\frac{p_0(1-p_1)}{p_1(1-p_0)}}$. This is known as the Neyman allocation.

(b) Using the Z-test statistic defined in part (a):

(i) Formulate an optimisation problem to find the allocation ratio that minimises expected failures subject to the power of the Z-test being constant as a function of parameters p_0, p_1.

(ii) Show that given a fixed sample size $n = n_0 + n_1$, such allocation ratio is equal to $R^* = \sqrt{\frac{p_0}{p_1}}\frac{(1-p_1)}{(1-p_0)}$.

[QUESTION CONTINUES ON THE NEXT PAGE]

(c) Now suppose the trial was conducted to simultaneously compare the new treatment ($k = 1$) and the standard treatment ($k = 0$) in J patient populations, indexed by $j = 1, \ldots, J$. For simplicity, assume a standard randomisation procedure with a fixed allocation ratio $R_j = 1$ was used for each substudy j. Focus on the hierarchical model below for a joint analysis of the binomial data from both substudies.

Let n_{jk} be the total number of patients on treatment $k = 0, 1$, within substudy $j = 1, \ldots, J$, and S_{jk} be the number of responses accordingly.

$$\begin{aligned} S_{jk} \mid n_{jk}, p_{jk} &\sim \text{Binomial}(n_{jk}, p_{jk}) \\ \log\left(\frac{p_{jk}}{1-p_{jk}}\right) &= \theta_{0j} + \theta_{1j} X_{jk}, \\ \theta_{1j} \mid \mu, \sigma &\sim N(\mu, \sigma^2), \qquad \text{for } j = 1, \ldots, J. \end{aligned}$$

Here, X_{jk} is an indicator for group $k = 0, 1$, in substudy $j = 1, \ldots, J$. We stipulate $X_{jk} = k$ so that $X_{j1} = 1$ indicates assignment to the experimental treatment and $X_{j0} = 0$ indicates the control. In the last line of the hierarchical model, μ and σ^2 are unknown parameters.

(i) Obtain point estimators in terms of μ and σ^2, using the principle of maximum likelihood;

(ii) State in words how to further find the maximum likelihood estimate for the response rates, p_{jk}, given the estimates for μ and σ^2.

Note: You should state the second order condition for questions (a) – (c), but do not need to verify it to get full marks.

3 Statistics in Medical Practice

A set of people are tested for a disease at weekly intervals. The disease can be fatal, but people can recover from it. If a person dies from the disease, their time of death is reported. We want to estimate the risk of getting the disease, the typical length of time spent with the disease, and the risk of death from the disease. We also want to estimate how the risk of getting the disease varies between people with and without a certain risk factor.

We assume the risk factor does not affect outcomes for people who have the disease, and we assume there is no risk of death from other causes. Assume also that the diagnostic test is completely accurate.

(a) Define a continuous-time, time-homogeneous multi-state model that can be used to answer these questions, indicating all unknown parameters, and briefly stating the meaning of each quantity you define.

(b) We observe a set of data which includes the following outcomes for two people, each of whom is disease-free at the start of the study.

- Person 1, doesn't have the risk factor: positive test 1 week after the start of the study, and a negative test 2 weeks after the start.
- Person 2, has the risk factor: died from the disease 6 days after the start of the study.

Define the contribution to the likelihood of the model in (a) for each of these people. All new symbols used must be defined, and it should be clear what parameters the likelihood is a function of. There is no need to express transition probabilities in terms of the model parameters.

(c) We are given the following estimates from the model:

(1) the mean duration of a single episode of illness is 10 days,

(2) the chance that an episode of the illness will be fatal is 10%,

(3) the chance that a person gets the illness at all within 30 days is 6% for people with the risk factor, and 1% for people without.

From these results, derive estimates of the transition rates that define the model in (a). Logarithms may be used without evaluation.

(d) Explain how we would then estimate the total amount of time that somebody who does not have the illness at the start of the study is expected to spend ill in the future, for a person with the risk factor and a person without the risk factor. Any formulae stated need not be evaluated.

(e) In parts (a)–(d) we supposed that the disease can only be detected through weekly testing. Suppose that in addition, when patients experience symptoms that may indicate having the disease, they report the date when those symptoms started to the study investigators.

Explain two different ways in which we might include this information in our analysis of disease risks. You do not have to write out likelihood functions, but you should briefly describe each proposed approach in words, stating any assumptions that it would be making.

4 Analysis of Survival Data

(a) A time-to-event dataset comprises n observations (x_i, v_i), $i = 1, \ldots, n$ where x_i is the time of event ($v_i = 1$) or censoring ($v_i = 0$), for the ith individual. There are no ties in the dataset, and the x_i are ordered such that $x_i > x_{i'}$ for $i > i'$.

(i) Explain what is meant by the *risk set* at time x_i. How many individuals are in the risk set at time x_i?

(ii) Assuming all individuals are exposed to the same hazard function, derive the *Nelson-Aalen* estimator of the integrated hazard function. Express your answer as a sum over the individuals in the dataset.

(b) What is a *Martingale residual*? Describe how to use Martingale residuals to assess whether a continuous explanatory variable z should be included in a time-to-event model.

(c) (i) Write down the Martingale residual for the ith individual in the dataset defined in part (a), where the common integrated hazard has been estimated using the Nelson-Aalen estimator.

(ii) Show that the sum of these residuals is equal to zero.

(d) Suppose that the dataset also includes z_i ($i = 1, \ldots, n$), the value of a continuous explanatory variable for the ith individual. A proportional hazards model is fitted to this data with baseline hazard $h_0(t)$ and hazard multiplier $\exp(\beta z_i)$ for the ith individual, where β is a scalar parameter.

(i) Write down the Martingale residual for the ith individual after fitting this model.

(ii) How would you interpret a Martingale residual with value 0.99? How would you interpret a Martingale residual with value -4?

Suppose further that there is a time c such that $v_i = 0$ for $x_i > c$.

[For the rest of this question, you may assume that: β is positive, z is a strong predictor of time-to-event, a substantial number of individuals have an event before time c, and a substantial number of individuals are censored before time c.]

(iii) What can you say about the minimum and maximum possible values of the Martingale residual corresponding to (A) an observed event and (B) a censored observation?

(iv) What, therefore, can you say about the distribution of points in a plot of the Martingale residuals against z? *[Your answer may be in the form of a sketch.]*

5 Analysis of Survival Data

A researcher enrols six patients into a study of time-to-death after starting a new treatment for their disease. A Kaplan-Meier estimate of the survivor function is calculated every 12 months after the start of the study.

All patients had died by 72 months after the study start. The following table shows the data available 48 months after the study start.

Patient	Month of Starting Treatment	Status at End of Month 48	If Dead, Month of Death
A	1	dead	26
B	4	dead	10
C	7	dead	19
D	12	alive	
E	13	dead	19
F	16	dead	42

All months in the table refer to months since study start. Treatments start at the beginning of a month, deaths occur at the end of a month. A patient starting treatment at the beginning of month p has been receiving treatment for *precisely* $q - p + 1$ months at the end of month q.

(i) For each patient write down the time from starting treatment to death or censoring (as appropriate), at 48 months after study start.

Let $\hat{\mathrm{F}}_M(t)$ denote the Kaplan-Meier estimate of the survivor function calculated using the data available at the end of month M after study start, with $M \in \{12, 24, 36, \dots\}$.

(ii) Calculate $\hat{\mathrm{F}}_{48}(t)$ for $t \geqslant 0$.

(iii) Calculate $\hat{\mathrm{F}}_{24}(t)$ for $t \geqslant 0$.

(iv) Why can we write down the value of $\hat{\mathrm{F}}_{36}(48)$ without performing the full Kaplan-Meier calculation? What can we say about $\hat{\mathrm{F}}_{48}(48)$?

(v) Calculate $\hat{\mathrm{F}}_{12}(12)$ and $\hat{\mathrm{F}}_{36}(12)$.

Let $\hat{\mathrm{F}}_\infty(t)$ be the final estimate of the survivor function in the sense that if $\hat{\mathrm{F}}_M(t) = \hat{\mathrm{F}}_{M^*}(t)$ for all $M > M^*$ then $\hat{\mathrm{F}}_\infty(t) = \hat{\mathrm{F}}_{M^*}(t)$. $\hat{\mathrm{F}}_\infty(t)$ is said to be *evaluable* at the M-monthly analysis if it is known at month M that there is sufficient data available to calculate $\hat{\mathrm{F}}_\infty(t)$.

(vi) What is $\hat{\mathrm{F}}_\infty(12)$? Which twelve-monthly analysis is the first at which $\hat{\mathrm{F}}_M(12) = \hat{\mathrm{F}}_\infty(12)$? Which twelve-monthly analysis is the earliest that $\hat{\mathrm{F}}_\infty(12)$ is evaluable?

(vii) Which twelve-monthly analysis is the earliest that $\hat{\mathrm{F}}_\infty(48)$ is evaluable?

[QUESTION CONTINUES ON THE NEXT PAGE]

At month 30, the researcher plans to submit the month 24 analysis to a journal.

(viii) What condition on the censoring mechanism is necessary for a valid Kaplan-Meier analysis? Is this condition satisfied at the 24-month analysis?

(ix) During the writing of the paper, the spouse of patient A tells the researcher that patient A died at the end of month 26 since study start. Should the researcher update the Kaplan-Meier analysis to take account of this additional information? Briefly justify your answer.

(x) If the researcher does update the Kaplan-Meier analysis, would that change your answer to part (viii)?

6 Analysis of Survival Data

(a) (i) What is a *competing risks* model in the context of time-to-event analysis?

(ii) What is meant by a *cause-specific* hazard?

(iii) Define the *cumulative incidence function*.

Suppose there are just two possible events A and B:

(iv) Derive the cumulative incidence function for event A in the presence of B, in terms of the cause-specific hazard functions.

Suppose further that the cause-specific hazards for A and B are constant, θ and ϕ respectively:

(v) What is the cumulative incidence function for A in the presence of B?

(vi) What is the cumulative incidence function for the composite event A or B?

[For the remainder of this question event A is to be interpreted as the event of interest and event B is to be interpreted as censoring.]

(b) A researcher is conducting a study in which the event of interest is the first occurrence of a side-effect (event A) after a patient has started taking a new medicine. Patients may withdraw from the study for reasons unrelated to the medicine (uninformative censoring: event B). There are n patients recruited into the study.

Assuming the cause-specific hazards for events A and B are θ and ϕ respectively:

(i) What is the expected number of patients who will be observed to experience event A?

(ii) What is the expected total length of time spent on study over all the patients (from starting to take the medicine to event A or censoring)?

(iii) What is the ratio of the expected number of patients to the expected total length of time?

The researcher decides, for safety reasons, that patients who have received the new treatment for c years should stop being treated, and their time-to-event should be censored at that point.

(iv) Repeat the calculations of parts (b)(i), (b)(ii) and (b)(iii) accounting for this additional source of censoring.

[QUESTION CONTINUES ON THE NEXT PAGE]

The dataset generated by this study has form $\{(x_i, v_i)\colon i = 1, \ldots, n\}$ with x_i being either the time of the first side-effect (event A, $v_i = 1$) or the time of censoring (event B or stopping treatment at c years, $v_i = 0$) for the ith individual.

(v) Derive the maximum likelihood estimator of θ in terms of v_+ (the total number of first side-effects observed) and x_+ (the total time at risk of a first side-effect). Comment on the relationship between the form of this estimator and your answer to part (b)(iv).

(vi) Derive and interpret the second derivative of the log-likelihood.

The researcher is particularly interested in θ near θ_0, and expects from previous experience the cause-specific hazard for censoring to be ϕ_0. The researcher intends to recruit sufficient patients into the study to satisfy:

$$\frac{\mathbb{E}[v_+|\theta = \theta_0, \phi = \phi_0]}{(\theta_0)^2} \geqslant I_0$$

where I_0 is a pre-specified quantity.

(vii) Why is this a reasonable approach to determining n? How should I_0 be chosen?

(viii) Use the results of part (b)(iv) to obtain n in terms of θ_0, ϕ_0, c and I_0.

END OF PAPER