UNIVERSITY OF CAMBRIDGE

MAMA/207, NST3AS/207, MAAS/207

MAT3

MATHEMATICAL TRIPOS Part III

Monday 9 June 2025 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

In the UK during World War 2, many foods were rationed; that is, people could only purchase a limited amount of that food. Rationing of sugar continued until September 1953. Figure 1 shows a striking increase in sugar consumption following the end of rationing.

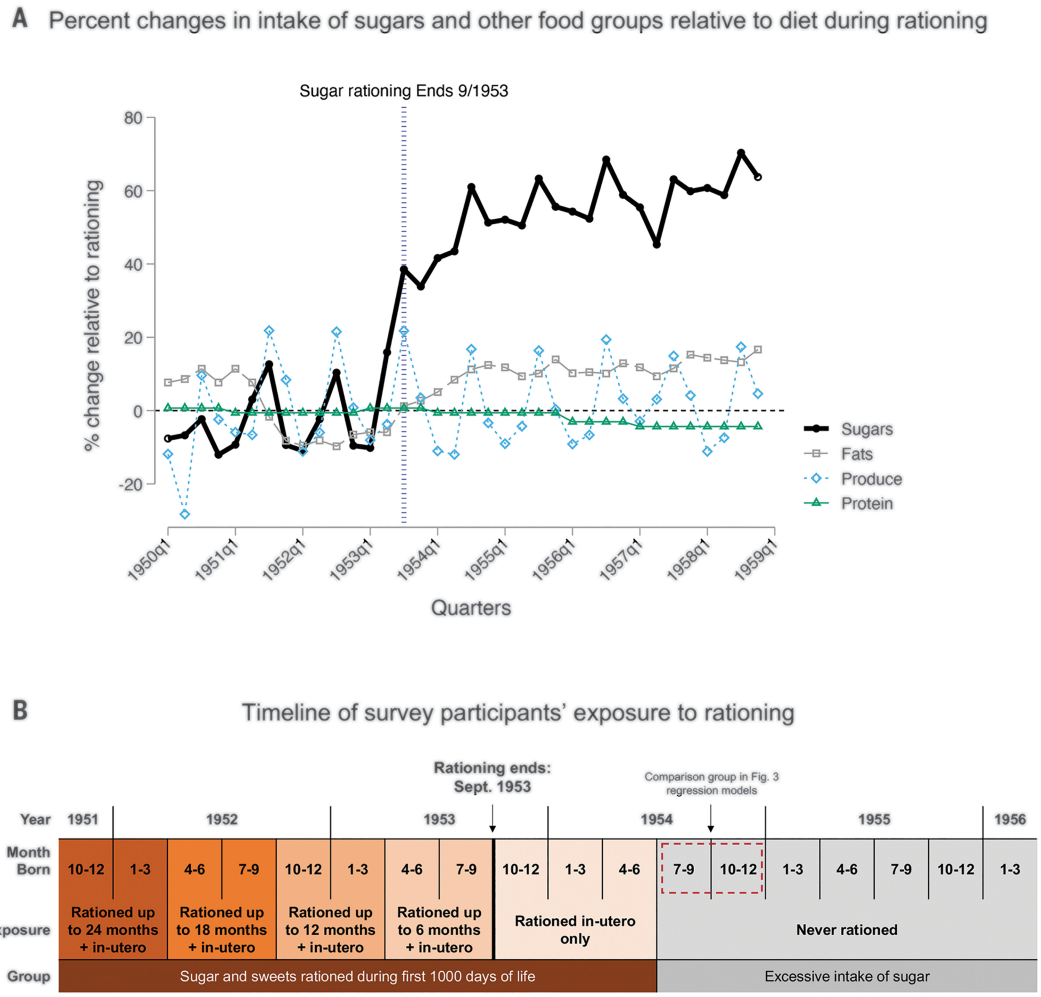

Figure 1. *Panel A*: food consumption during quarters of the calendar year. 1950q1 indicates the first quarter of year 1950. *Panel B*: subgroups of the population divided by quarter of birth year. Taken from Gracner et al., Science 2024.

Suppose we are interested in testing whether sugar consumption has a causal effect on Type 2 diabetes risk in later life. We have data on participants from UK Biobank, a cross-sectional sample of the UK population.

[QUESTION CONTINUES ON THE NEXT PAGE]

(a) Using counterfactual language, write down the null and alternative hypotheses that we wish to test.

(b) What assumptions do we need to make for comparison of those born in different periods to be a valid test of the causal null hypothesis?

(c) Give three reasons why an association between birth period and Type 2 diabetes risk may occur even if there is no causal effect of sugar rationing on Type 2 diabetes risk. In each case, describe how you could design the analysis to minimize this possibility.

(d) Suppose there is a subset of the population who abstain from eating sugar for cultural reasons. How could analysis of this group strengthen our causal claim?

(e) The UK Biobank study is not fully representative of the UK population. In particular, healthy and affluent people are overrepresented. How could this affect the interpretation of our findings in terms of: (i) internal validity, and (ii) external validity (i.e. generalizability of findings).

2 Statistics in Medical Practice

We observe a dataset where people are tested for an infection at the start of every month. We wish to estimate the expected length of an episode of infection, and the expected time from the end of this episode until a person's next infection. Assume that getting the infection does not give immunity, that the test for the infection is perfect, and that individuals do not interact.

(a) Describe a multistate model that might be used to answer this question, indicating the parameters, and explain how the parameters are related to the quantities we want to estimate.

(b) Suppose we hypothesise that an infection lasts 1.25 months on average, and that there is an average of 5 months until the next infection. A person provides the following data:

Month 1	test negative
Month 2	test negative
Month 3	test positive

Prove that this person's contribution to the likelihood of the multistate model, given the parameter values implied by our hypothesis, is

$$0.16 - 0.12\exp(-1) - 0.04\exp(-2)$$

You may use, without proof, the result that the matrix exponential of

$$\begin{pmatrix} -a & a \\ b & -b \end{pmatrix} t$$

has an entry of $\left(b + ae^{-(a+b)t}\right)$ *in the first row and first column.*

Suppose now that we want to acknowledge that getting an infection increases a person's immunity from subsequent infections.

(c) What does this judgement imply about the parameters of the model in parts (a,b), and why would a model extended in this way be difficult to implement given the data that we have observed?

[QUESTION CONTINUES ON THE NEXT PAGE]

Now instead, suppose that we are modelling a population of individuals who may transmit the infection between each other, and the aim is to estimate the rate of incidence of infection.

(d) Draw diagrams indicating the states and transitions for:

(i) the basic SIR model

(ii) as in (i), but with a pre-infectious period of latent infection

(iii) as in (i), but with non-lasting immunity.

In each of the three models, label all arcs with an appropriately defined transition rate, and write down an expression for d, the expected duration of an infection. Denote the time-t infection hazard by $\lambda(t)$.

(e) Write $\lambda(t)$ in terms of a transmission rate parameter β, and appropriate variables for the number of people in each state and population size, under the laws of mass action. Use these to formulate an expression for the basic reproduction number.

(f) Typically, infections are not directly observed. Instead, time series of some sequelae of infection might be available, such as new admissions to hospital among infected individuals. Assume a continuous-time formulation of a model, where a fraction, p of infections require hospital admission, and assume that hospital admission occurs at a time post-infection that is distributed according to a probability distribution f. Write down an equation that links p, $f()$ and the infection hazard, $\lambda(t)$, to $\mu(t)$, the incident rate at which new admissions occur.

(g) Assume that $f(t) = \sigma e^{-\sigma t}$, the pdf of an exponential distribution. Use integration by parts to show that

$$\frac{d\mu}{dt}(t) = \sigma\left(p\lambda(t) - \mu(t)\right),$$

adding a brief comment on the interpretation of this result.

3 Statistics in Medical Practice

In a two-arm parallel-group trial, let π_1 and π_0 respectively denote the probability of treatment response in the treatment and control groups, and let X_1 and X_0 represent the number of successes out of n_1 and n_0 observations in the two groups, so that the probability of success in the treatment and control groups can be estimated by $p_1 = X_1/n_1$ and $p_0 = X_0/n_0$, respectively, where both groups are independent and where X_1 and X_0 are assumed to follow binomial distributions:

$$b(n_1, \pi_1, x) = \mathbb{P}(X_1 = x) = \binom{n_1}{x} \pi_1^x (1-\pi_1)^{n_1 - x}$$
$$b(n_0, \pi_0, x) = \mathbb{P}(X_0 = x) = \binom{n_0}{x} \pi_0^x (1-\pi_0)^{n_0 - x}$$

with $\mathrm{E}[X_1] = n_1\pi_1$ and $Var[X_1] = n_1\pi_1(1-\pi_1)$.

We are interested in testing the hypothesis $H_0 : \delta \leqslant 0$ versus $H_1 : \delta > 0$, where $\delta = \pi_1 - \pi_0$ denotes the target difference of interest.

(a) Noting that

$$\bar{p} = \frac{X_1 + X_0}{n_1 + n_0} = p_1 \left(\frac{n_1}{n_1 + n_0} \right) + p_0 \left(\frac{n_0}{n_1 + n_0} \right)$$

is the best estimate of π_1 and π_0 under the null hypothesis and that

$$\bar{p} \approx \pi_1 \left(\frac{n_1}{n_1 + n_0} \right) + \pi_0 \left(\frac{n_0}{n_1 + n_0} \right) = \bar{\pi},$$

write down a suitable Wald T test statistic for testing H_0, expressing its denominator as a function of $\bar{\pi}$, n_1 and n_0.

(b) Obtain $\mathrm{E}[T]$ and $\mathrm{Var}[T]$, the expected mean and variance of the test statistic defined in (a), hence state the asymptotic distribution of T:

(i) under H_0,

(ii) under the alternative hypothesis $\delta = \delta_A$.

(c) Obtain a formula for the sample size required to detect a difference of δ_A with a power of $1 - \beta$ (where β denotes the type II error rate) using a one-sided test with a type I error rate of α and assuming equal allocation to both groups (i.e., $n_1 = n_0$).

[QUESTION CONTINUES ON THE NEXT PAGE]

Suppose now that we wish to design a single-arm trial to learn more about a new treatment for pulmonary arterial hypertension. Specifically, we wish to determine if the (binary) response rate for this new treatment is greater than for the current standard of care.

(d) State the null hypothesis that will be tested by this trial, precisely defining all quantities in your answer.

(e) A Mander & Thompson two-stage design is being considered for this trial, with the following design parameters: $\{n_1 = 22, N = 54, r_1 = 7, e_1 = 10, r = 20\}$, representing (respectively) sample size at interim analysis, maximum sample size, stopping bounds at interim analysis and final stopping bound.

Derive an equation for the type-I error-rate of this particular realisation/instance of the Mander & Thompson design, in terms of response rate $\pi_0 = 0.3$ and the given design parameters, and using the probability mass function of the binomial distribution given in the introduction.

4 Analysis of Survival Data

Explain the principles of the *log-rank* test for comparing two survival distributions. When is it most powerful? Give an example of circumstances in which the two survival distributions are quite different but the log-rank test is unlikely to detect the difference.

Patients receiving an experimental treatment for cancer are assessed every month to determine whether the treatment has failed to stop the growth of the cancer. The following (artificial) data shows the time in months from starting the treatment to treatment failure of the 10 different patients taking part in the trial.

Group A $1, 2, 3+, 3+, 5$
Group B $1+, 5, 5, 7+, 9+$

where a plus sign indicates a right-censored value. Describe briefly any difficulties which may result from data being collected at fixed timepoints.

The following two scenarios refer to this dataset.

(a) Suppose that Group A comprises those patients randomised to a low dose of the experimental treatment and Group B comprises those randomised to a high dose.

(i) Calculate the expected number of treatment failures under the low dose, under the null hypothesis of no difference between the low and high dose.

(ii) Under the null hypothesis of no difference between the two doses, is the expected number of treatment failures under the high dose equal to the expected number under the low dose. If not, why not?

(iii) Define and calculate the *log-rank statistic* (you need not normalise to unit variance). Interpret the sign of the statistic.

(iv) Calculate a measure of the low versus high dose *relative risk* of treatment failure.

(b) Suppose, instead, all patients receive the same dose of the experimental treatment. The patients may or may not receive an additional treatment (surgery), the decision being made 3 months after the start of treatment. Group B now comprises those patients deemed fit enough for surgery, and Group A comprises those deemed too ill for surgery.

Is it a appropriate to use a log-rank test to compare the two groups in this scenario? Briefly justify your answer.

5 Analysis of Survival Data

(a) Describe how to obtain a maximum likelihood estimator of the parameters of a fully parametric time-to-event distribution, in the presence of right censoring.

(b) Obtain the maximum likelihood estimator for λ when the hazard function for the ith individual is given by $\mathrm{h}_i(t) = \lambda$.

(c) Consider the case where the hazard function for the ith individual is given by:

$$\mathrm{h}_i(t) = \theta \exp(z_i \beta) \text{ with } z_i \in \{0, 1\} \ .$$

Derive the log-likelihood function for (θ, β) and hence obtain the maximum likelihood estimator $\hat{\beta}$ of β. (You need not verify that the stationary point that you obtain is a maximum.)

[Hint: you may find it helpful to split your summations over i thus: $\sum_i = \sum_{i:z_i=0} + \sum_{i:z_i=1}$ *.]*

(d) Divide the individuals into two groups $k \in \{0, 1\}$ such that group k includes individuals $\{i : z_i = k\}$. Let λ_k denote the hazard experienced by individuals in group k.

How are the parameters (λ_0, λ_1) related to (θ, β)? Write down expressions for $\hat{\lambda}_0$ and $\hat{\lambda}_1$ and verify that the relationship between (λ_0, λ_1) and (θ, β) still holds when maximum likelihood estimators are substituted for parameters.

6 Analysis of Survival Data

What is meant by a *frailty* model in time-to-event analysis?

A *proportional frailty* model formulates the conditional hazard as

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

where U ($U > 0$) is a frailty variable with density $g(u)$ and $h_0(t)$ is a finite baseline hazard. Obtain an expression for the unconditional hazard function $\bar{h}(t)$ in terms of g and $h_0(t)$. Derive a condition on the frailty distribution for $\bar{h}(0)$ to equal $h_0(0)$.

(a) For each of the following frailty densities calculate $\bar{h}(t)$ when $h_0(t) = \theta$, and comment on the values of $\bar{h}(0)$ and $\lim_{t\to\infty} \bar{h}(t)$.

(i)
$$g(u) = \frac{1}{2}\delta(u - 1/2) + \frac{1}{2}\delta(u - 3/2) \ .$$

(ii)
$$g(u) = \exp(-u) \ .$$

(b) How, in general, would you expect the expectation of U conditional on no event before t to vary with time?

Using the frailty density specified in part(a)(i) calculate $\gamma(u, t)$, the density of U conditional on no event before t, and verify that $\gamma(u, 0) = g(u)$. Calculate $\mathbb{E}[U|T \geqslant t]$ and comment on the dependence of this expectation on time.

(c) Explain how the existence of frailty can affect the interpretation of a proportional hazards model. Illustrate your answer by using the frailty density specified in part(a)(ii) to show that unconditional hazards may not be proportional even if conditional hazards are proportional.

END OF PAPER