UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 10 June, 2022 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

Vitamin D has been proposed as a potential causal determinant of mortality risk. Observational epidemiological studies have consistently found that low concentrations of circulating 25-hydroxyvitamin D [25(OH)D] (a substance found in blood that is used as a clinical indicator of vitamin D status) are associated with higher risk of all-cause mortality. However, several large randomized trials of vitamin D supplementation on mortality risk have reported null results.

1. What does it mean for vitamin D to be a "causal determinant of mortality risk"?
2. Why may observational studies and randomized trials provide conflicting results? Provide two reasons with justifications.

An efficient approach for assessing the potential causal effect of vitamin D supplementation is Mendelian randomization, the use of genetic variants as instrumental variables to assess evidence for the effect of an exposure on an outcome. In a recently published paper, the authors constructed two candidate instruments: a focused instrument and a polygenic instrument. The focused instrument is a weighted sum of genetic variants in four gene regions that are known to relate to vitamin D metabolism. The polygenic instrument is a weighted sum of 71 genetic variants that are associated with 25(OH)D levels, but their biological relevance to the exposure is less clear. Associations of the focused and polygenic instruments with various traits are shown in Figure 1.

3. How do the plots in Figure 1 help understand the validity of the candidate instruments? Which candidate instrumental variable would you prefer?

It is suspected that the causal effect of vitamin D supplementation is stronger for individuals with low levels of 25(OH)D. The authors considered stratifying the population on 25(OH)D levels and calculating instrumental variable estimates for strata of the population, by estimating associations of the focused instrument with 25(OH)D and mortality risk for individuals with 25(OH)D below 25 nmol/L, with 25(OH)D between 25 and 50 nmol/L, and so on.

4. Why could stratifying the population by conditioning on the exposure in this way lead to the instrumental variable assumptions being violated? It may help to draw a directed acyclic graph of the instrumental variable assumptions.

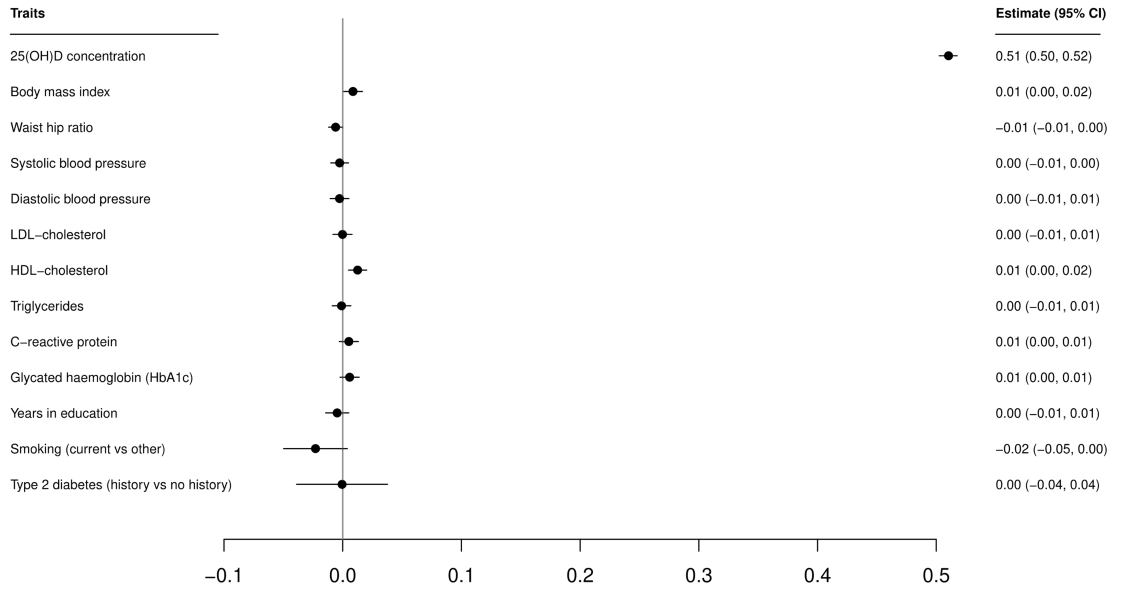
Instead, the authors first regressed 25(OH)D levels on the focused instrument, took the residual from this regression, centred the residual to have the same population mean as the original 25(OH)D measurement, and stratified on that variable, referred to as "residual 25(OH)D". They estimated associations of the focused instrument with 25(OH)D levels and with mortality risk in each stratum, and used these associations to calculate instrumental variable estimates, representing odds ratios for all-cause mortality scaled to a 10 nmol/L increase in 25(OH)D concentrations. Instrumental variable estimates in the population as a whole, and in each stratum of the population are shown in Figure 2.

5. Provide a brief explanation of the results of Figure 2. What do these results tell us about the potential causal effect of vitamin D supplementation?
6. The dataset used for this analysis had information on several other risk factors and disease outcomes. Provide and justify two ways that the authors could make their causal claims more convincing without collecting additional data.

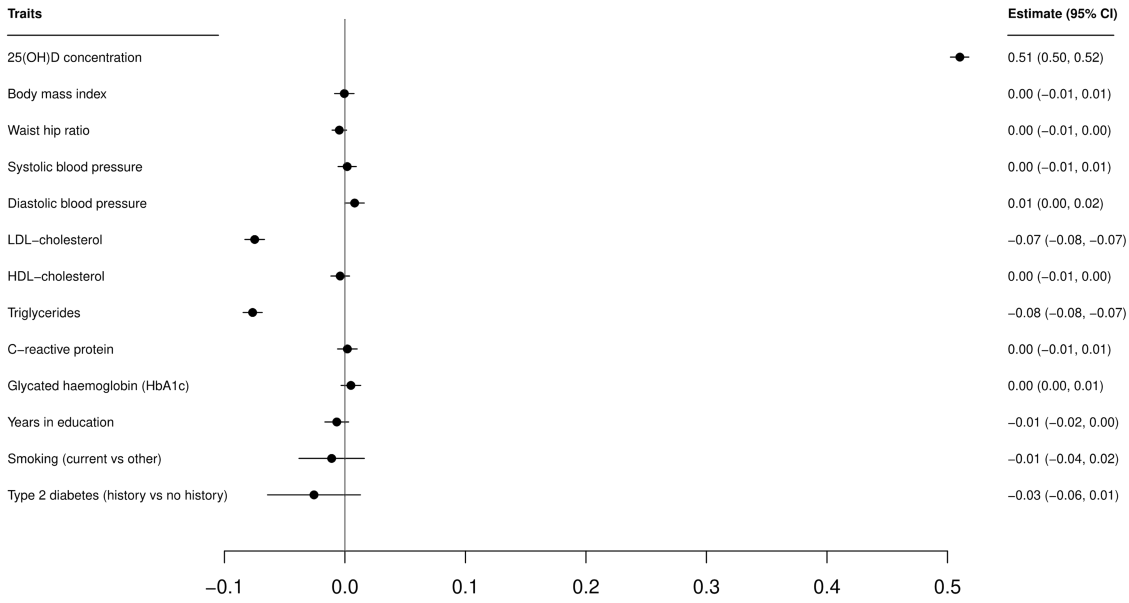
[QUESTION CONTINUES ON THE NEXT PAGE]

Figure 1. Estimates of associations of the candidate instruments with various traits. Error bars represent 95% confidence intervals (CI). Estimates for all continuous traits expressed in standard deviation units. Estimates for the binary traits (smoking and Type 2 diabetes status) are log odds ratios. Associations are scaled to a 10 nmol/L increase in genetically-predicted 25(OH)D concentrations.

Focused instrument:

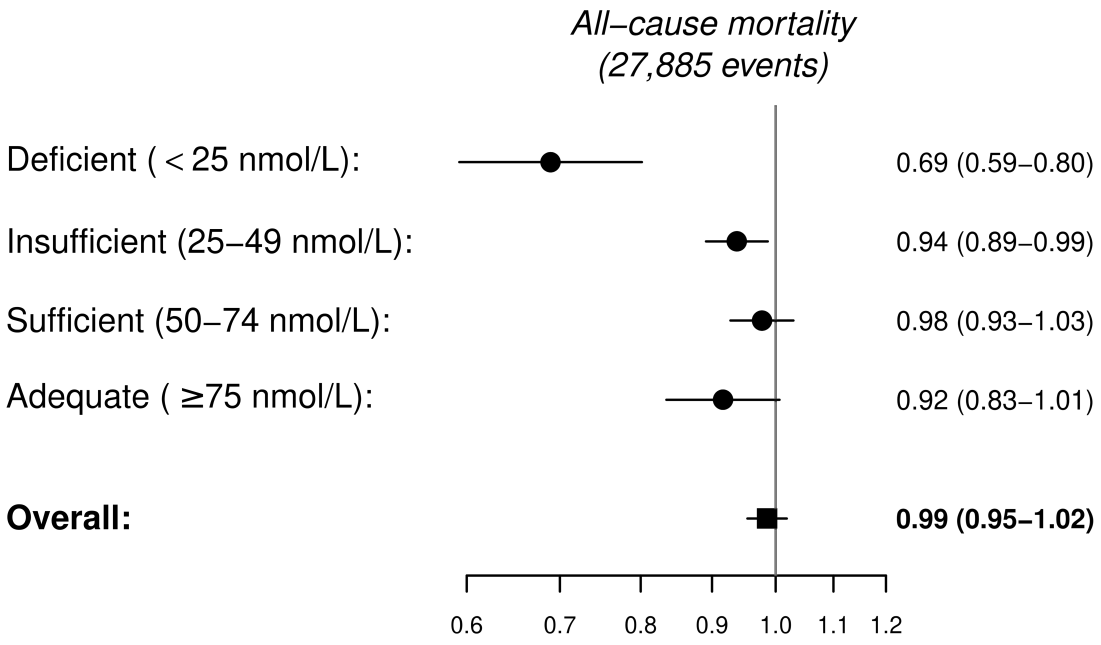


Polygenic instrument:



[QUESTION CONTINUES ON THE NEXT PAGE]

Figure 2. Mendelian randomization estimates for mortality in overall population and in strata defined by residual 25(OH)D concentrations. Estimates (95% confidence intervals) represent odds ratios per 10 nmol/L higher genetically-predicted concentration of 25(OH)D.



2 Statistics in Medical Practice

A clinical trial was carried out of a new treatment designed to control the symptoms of a particular disease. Patients with the disease were randomised with equal probability to receive either the new treatment or an existing treatment. One month after randomisation, patients were asked to attend a clinic to have their symptoms assessed and classified as either controlled (which is a treatment success) or uncontrolled (a treatment failure). Let $Z = 1$ for patients randomised to the new treatment and let $Z = 0$ for patients randomised to the existing treatment. Let $Y = 1$ for patients whose symptoms were controlled at one month after randomisation, and $Y = 0$ for patients whose symptoms were uncontrolled.

The following table shows the numbers of patients with controlled and uncontrolled symptoms in the two treatment groups.

Z	Y	Patients
0	0	80
0	1	120
1	0	110
1	1	110

Let $\phi_0 = P(Y = 1 \mid Z = 0)$ and $\phi_1 = P(Y = 1 \mid Z = 1)$. Let $\alpha = \phi_1 - \phi_0$ denote the treatment effect at one month after randomisation.

1. Write down the likelihood function for the observed data on these 420 patients in terms of ϕ_0 and ϕ_1.

2. Write down the maximum likelihood estimates (MLE) of ϕ_0 and ϕ_1. Verify that the MLE of α equals $-\frac{1}{10}$. (Note that you are not expected to derive these MLEs.)

In addition to the 420 patients represented in the table, there were 70 patients in the study who did not attend the clinic one month after randomisation, and so have a missing value of Y. Of these 70 patients, 50 had been randomised to the existing treatment and 20 had been randomised to the new treatment. The MLEs of ϕ_0, ϕ_1 and α that you have just calculated are called 'complete-case' MLEs, because they only use the data on the 420 'complete cases' (i.e. patients whose data on Z and Y are observed).

3. Briefly explain why the complete-case MLE of α may be biased.

4. Carefully derive the observed-data likelihood for the data on all 490 patients in terms of ϕ_0 and ϕ_1. Verify that the observed-data MLE of α (i.e. the MLE calculated from this likelihood) is equal to the complete-case MLE (i.e. $-\frac{1}{10}$).

[QUESTION CONTINUES ON THE NEXT PAGE]

The 420 patients who attended the clinic one month after randomisation were asked to attend the clinic again five months later, and 300 of these patients did so. The 70 patients who did not attend the clinic one month after randomisation were not asked to attend the clinic again. Let $W = 1$ for patients whose symptoms were controlled at six months after randomisation, and $W = 0$ for patients whose symptoms were uncontrolled. The following table shows the numbers of patients with controlled and uncontrolled symptoms at one month and six months after randomisation.

Z	Y	W	Patients
0	0	0	20
0	0	1	20
0	1	0	15
0	1	1	75
0	0	missing	40
0	1	missing	30
0	missing	missing	50
1	0	0	50
1	0	1	30
1	1	0	30
1	1	1	60
1	0	missing	30
1	1	missing	20
1	missing	missing	20

Let $\psi_0 = P(W = 1 \mid Z = 0)$ and $\psi_1 = P(W = 1 \mid Z = 1)$. Let $\beta = \psi_1 - \psi_0$ denote the treatment effect at six months after randomisation.

5. State whether the data on (Z, Y, W) are monotone missing.

6. Explain in terms of dropout what is meant by the assumption that the data on (Z, X, W) are missing at random.

7. Suppose that the data (Z, X, W) are assumed to be missing at random and that (improper) multiple imputation is carried out to estimate ϕ_0, ϕ_1, α, ψ_0, ψ_1 and β. Suppose this is done using the following (saturated) imputation models:

$$\begin{aligned} P(Y = 1 \mid Z) &= \gamma_{00} + \gamma_{01} Z \\ P(W = 1 \mid Z, Y) &= \gamma_{10} + \gamma_{11} Z + \gamma_{12} Y + \gamma_{13} ZY \end{aligned}$$

where $\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}, \gamma_{12}$ and γ_{13} are unknown parameters. It can be shown that the resulting estimators of ϕ_1 and ψ_1 converge to the values $\frac{1}{2}$ and $\frac{25}{48}$, respectively, as the number of imputed datasets tends to infinity.

   (a) Calculate the values to which the estimators of ϕ_0 and α converge.

   (b) Calculate the values to which the estimators of ψ_0 and β converge.

3 Statistics in Medical Practice

An infectious disease is seeded in a closed population of size N, and is such that infected individuals are immediately and forever infectious.

(a) Draw a simple compartmental structure representing the transmission process of this disease, using an appropriately defined transmission rate parameter β.

(b) Write down a system of ordinary differential equations governing the transmission dynamics of this disease, including the support of any parameters.

(c) Assume that initially, there are $\frac{\theta}{\theta+1}N$ susceptible individuals and $\frac{1}{\theta+1}N$ individuals infectious with the virus. Show that the rate of change in the number of infectious individuals, $\Lambda(t)$, can be written as a function of time by:

$$\Lambda(t) = \frac{\theta N \beta e^{-\beta t}}{(1+\theta e^{-\beta t})^2}$$

[Note: Depending on how you have answered (a) a scaling of the β parameter may be required.]

(d) Obtain the value of $\Lambda(t)$ at its maximum.

(e) Simplify the equation for $\Lambda(t)$, using the time at the maximum as the origin. Comment briefly on the shape of the curve traced out by $\Lambda(t)$.

The dynamics of an epidemic with removal, such as HIV in the anti-retroviral treatment era, in which sufficient treatment can lead to viral suppression (and hence removal of a patient from the infectious population), can be represented by a SIR model given by the equations

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta S(t) I(t) \\
\frac{dI(t)}{dt} &= \beta S(t) I(t) - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t),
\end{aligned}$$

for $\beta, \gamma > 0$. Assume that initially there are no removed individuals and that initially the number of infectious individuals is very small, among the total (closed) population of size N.

(f) Show that the peak number of removals per unit time occurs when $S(t) = \gamma/\beta$.

(g) Evaluate I and R at the time of the peak in removals, in terms of the initial number of susceptible individuals and the basic reproduction ratio R_0.

(h) Show that a relation for the final size of the epidemic, $R(\infty)$, in terms of R_0 is

$$R(\infty) \approx \frac{S(0)}{R_0} \log \frac{S(0)}{S(0) - R(\infty)}$$

4 Analysis of Survival Data

Derive the *log-rank* test for comparing two time-to-event distributions. (You may use the result that if n individuals are randomly allocated to the four cells of a two-by-two contingency table, such that (i) the row totals p_1, p_2 and the column totals q_1, q_2 are fixed and (ii) the probability of an individual being allocated to a particular row does not depend on the column to which that individual has been allocated, then - defining U as the number of individuals allocated to the cell in the first row and first column - U has expectation $p_1 q_1/n$ and variance $p_1 p_2 q_1 q_2/n^2/(n-1)$.)

The table shows the survival in months of ten different patients after two different treatments for cancer (artificial data):

Treatment	Survival Time
A	1, 2, 3+, 3+, 5
B	1+, 5, 5, 7+, 10+

where a plus sign indicates a right censored value.

The following parts of the question refer to this dataset.

(a) Calculate the expected number of observed deaths under Treatment A, under the null hypothesis of no difference in time-to-death distribution between Treatments A and B. Why is this expected number not half of the total number of deaths?

Calculate a measure of the *relative risk* of Treatment A relative to Treatment B.

(b) Write down the *maximum likelihood* estimator for the hazard when the time-to-event distribution is exponential.

Assuming the time-to-event distribution is exponential for both Treatments A and B, show that the maximum likelihood estimate of the hazard ratio (Treatment A relative to Treatment B) is equal to 3.

(c) Write down the *Nelson-Aalen* estimator for the integrated hazard function, assuming that all event times are distinct. Suggest a modification for the estimator when there are tied event times.

Calculate the ratio of the Nelson-Aalen estimates of the integrated hazards (Treatment A relative to Treatment B) at month 5.

(d) How do the three ratios calculated in parts (a), (b) and (c) compare with one another?

5 Analysis of Survival Data

(a) The independent continuous time-to-event random variables T_i for two individuals $i \in \{1, 2\}$ have densities $\mathrm{f}_i(t)$ and survivor functions $\mathrm{F}_i(t)$.

(i) Show that, in the absence of censoring, the probability that individual 1 has an event before individual 2 is given by:

$$\mathbb{P}[T_1 < T_2] = \int_0^\infty \mathrm{f}_1(t')\mathrm{F}_2(t')\,\mathrm{d}t' .$$

(ii) Suppose that both individuals are subject to a fixed censoring time c. The event times t_1, t_2 are described as *informative* if it is possible to determine which individual was first to have an event. Show that the pair t_1, t_2 is informative if $t_1 \neq t_2$ and $\min(t_1, t_2) \leqslant c$.

Find the probability that $T_1 < T_2$ given that the pair of event times are informative.

(b) Consider the case when T_1, T_2 have exponential distributions with rate parameters λ_1, λ_2 respectively.

(i) Show that, when the two individuals are subject to a common fixed censoring time c:

$$\mathbb{P}[T_1 < T_2 | T_1, T_2 \text{ are informative}] = \frac{\lambda_1}{\lambda_1 + \lambda_2} .$$

(ii) Explain why this result generalises to the two individuals each being subject to a random censoring time distribution.

(c) Consider now the case when the ith individual is subject to integrated hazard $\phi_i H_0(t)$ where the ϕ_i are constants and H_0 is a baseline integrated hazard function with inverse H_0^{-1}.

(i) Define time-to-event variable U_i by $U_i = H_0(T_i)$. What is the distribution of U_i?

(ii) Deduce that:

$$\mathbb{P}[T_1 < T_2 | T_1, T_2 \text{ are informative}] = \frac{\phi_1}{\phi_1 + \phi_2} .$$

when the two individuals each are subject to a random censoring time distribution. c.

(d) What is the connection between this question and competing risks methodology?

6 Analysis of Survival Data

What is a *proportional frailty* model? Consider a proportional frailty model with frailty variable U and baseline hazard $h_0(t)$. Obtain an expression for the unconditional (population) survivor function in terms of the density $g(u)$ of U and the baseline integrated hazard. Show that, without loss of generality, if U has a finite mean then U can be defined such that $\mathbb{E}U = 1$. Why would you want to do so?

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

(a) Derive an expression for the unconditional hazard $\bar{\mathrm{h}}(t)$ as a function of the baseline integrated hazard.

(b) Derive the density $g(u, t)$ of U at time t over individuals who have not had an event by t.

(c) What is the expectation of U at time t over individuals who have not had an event by t? Comment on how this expectation changes with t.

(d) Obtain $\bar{\mathrm{h}}(t)$ directly from your answer to part (c).

END OF PAPER