The University Of Sheffield.

PLEASE LEAVE THIS EXAM PAPER ON YOUR DESK.
DO NOT REMOVE IT FROM THE HALL.

Data Provided:
Neaves Tables
Graph Paper

SCHOOL OF MATHEMATICS AND STATISTICS MAS6061

Session 2016-2017 **3 Hours**

Epidemiology and Time Series

RESTRICTED OPEN BOOK EXAMINATION.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.
*All answers will be marked but credit will be given for only the best **FIVE** answers.*
All questions carry equal marks. Total marks 100.

Registration number from U-Card (9 digits) – to be completed by student

(This page is left blank)

1. The data in the table below describe an unmatched case-control study to determine the effect of bisphosphonate use on atrial fibrillation. The cases were 3141 women diagnosed at a Danish hospital with atrial fibrillation and the controls were 4269 women selected from population registers. Use of bisphosphonates in cases and controls was identified using population based prescription databases.

Age (Years)	Atrial Fibrillation	Bisphosphonate Use	
		Exposed	Not Exposed
≤70	Cases	47	816
	Controls	55	1053
71-80	Cases	120	901
	Controls	69	1417
>80	Cases	72	1185
	Controls	95	1580

We are interested in determining the effect of bisphosphonate use on the risk of atrial fibrillation.

(i) Explain why we cannot estimate the relative risk directly in this study? **(2 marks)**

(ii) Ignoring age, calculate the odds ratio and 95% confidence interval for diagnosis of atrial fibrillation for women who have taken bisphosphates relative to women who have never taken bisphosphonates. Comment on the results. **(6 marks)**

(iii) Use an appropriate method to calculate an estimate of the odds ratio for diagnosis of atrial fibrillation in women due to bisphosphonate use, allowing for age. Calculate a 95% confidence interval for this odds ratio. **(9 marks)**

(iv) Compare the results from (ii) and (iii). Is there evidence of a significant association between bisphosphonate use and atrial fibrillation? **(3 marks)**

2. Rheumatoid arthritis (RA) is an autoimmune disease with a strong genetic component. A genome-wide association study of RA severity has found a statistically significant association with a genetic variant CNV45.
A Sheffield study has genotyped a cohort of 245 RA patients for this variant. 132 of these patients had a mild form of disease. They also genotyped the variant in 398 local female blood donor controls. The results of the study are summarised below. The Hardy Weinberg Equilibrium (HWE) test has been reported for each row of data in the Table.

		CNV45 genotype				
		AA	AT	TT	Row Total	HWE p-value
RA Patients	Mild form	25	77	30	132	0.053
	Severe form	17	52	44	113	0.800
Local blood donor controls		200	160	38	398	0.122

(i) What is Hardy Weinberg Equilibrium (HWE) and why is the HWE test often reported in candidate gene association studies? **(2 marks)**

(ii) How do you interpret the HWE results reported here and are they all relevant? **(2 marks)**

(iii) Use an appropriate statistical hypothesis test to compare the allele frequencies of the CNV45 alleles in severe RA patients to mild RA patients. Is there any evidence that the CNV45 variant is associated with RA severity? **(4 marks**)

(iv) Report appropriate comparative risk statistics for the risk of severe RA (compared to risk of mild RA) for the exposure levels of genotype. Compare these two comparative risk statistics (without calculating 95% confidence intervals) and comment whether the CNV45 allele A appears to act in a dominant/codominant or recessive manner to allele T. **(4 marks)**

(v) Use an appropriate statistical hypothesis test to evaluate the evidence that the CNV45 allele is associated with risk of rheumatoid arthritis. **(4 marks)**

2 continued

(vi) The researchers are concerned that duration of disease may be confounding this potential association. They tried to address this by matching severe and mild cases by duration of disease, which led to 106 pairs summarised in the table below.

	Severe case AA	Severe case AT	Severe case TT	Total
Mild case AA	16	5	3	24
Mild case AT	0	40	30	70
Mild case TT	1	1	10	12
Total	17	46	43	106

Discuss how this paired data could be used to assess the effect of the genotypes on severity of RA whilst controlling for duration of disease. Use conclusions drawn in iv) to simplify the data table and perform an appropriate hypothesis test. Comment on the results. **(4 marks)**

3. A large population study is conducted to assess the association between obesity and cardiovascular disease in secondary school teachers. Teachers aged between 35 and 55 were recruited to the study and followed up for ten years. In this study obesity was defined as BMI greater than or equal to 30 at baseline recruitment. Cardiovascular disease (CVD) risk is known to increase with age, so age may be an important confounding variable. The results of the study are shown below.

	Age <= 45		**Age > 45**	
	CVD	No CVD	CVD	No CVD
BMI >= 30	105	450	240	600
BMI < 30	120	1800	60	900
	225	2250	300	1500

(i) According to the classic definition of confounding, is there evidence of confounding in the relationship between BMI and being diagnosed with CVD?
(4 marks)

(ii) Calculate the difference in risks *(Risk Difference)* for cardiovascular disease by BMI exposure for the subgroup aged 45 and below at study onset, the subgroup aged over 45 and the collapsed (over age group) table i.e. three risk differences in total. **(6 marks)**

(iii) Comment on whether there is evidence of confounding according to the collapsibility definition. **(2 marks)**

(iv) Calculate the crude and standardised expected counts to assess if age is a confounding variable according to the counterfactual definition. Report the standardised risk difference. **(4 marks)**

(v) Compare your answers from sections i) to vi). What is the overall evidence that age is a confounder and/or an effect modifier in the relationship between BMI and CVD in secondary school teachers? How are the results of this association best reported (i.e. should crude, adjusted or strata specific risk differences be reported)?
(4 marks)

4 (i) Explain briefly why the definition of the sample autocorrelation function is not appropriate for non-stationary time series. ***(2 marks)***

(ii) (a) For the following time series data

time (t)	response (y_t)
1	1
2	3
3	4
4	10
5	2
6	2
7	5
8	11

calculate 4-span moving averages for time points $t = 3, 4, 5, 6$. ***(4 marks)***

(b) Consider the time series

$$y_t = \alpha + \beta t,$$

for some known α and β and for $t = 1, 2, 3, \ldots$. Show that the 4-span moving average of $\{y_t\}$ at time t is exactly equal to the value of y_t. ***(4 marks)***

(c) Now consider the time series

$$x_t = y_t + \epsilon_t,$$

where y_t is the time series in part (b) above and ϵ_t is white noise with variance 2. Show that the de-trended time series defined by

$$d_t = x_t - \hat{y}_t,$$

where $\hat{y}_t$ is the 4-span moving average of y_t, is a weakly stationary time series. ***(2 marks)***

(iii) Let $\{Z_t\}$ be a sequence of independent identically distributed random variables, so that Z_t follows a normal distribution with zero mean and variance 1.

Define the process

$$y_t = \begin{cases} Z_t, & \text{if } t \text{ is even} \\ \dfrac{Z_t^2 - 1}{\sqrt{2}}, & \text{if } t \text{ is odd.} \end{cases}$$

(a) Show that $\{y_t\}$ is a white noise process $WN(0, 1)$. ***(6 marks)***

(b) Show $\{y_t\}$ is not an identically distributed sequence. ***(2 marks)***

HINT: If a random variable X follows the chi-square distribution with ν degrees of freedom $X \sim \chi^2_\nu$, then $E(X) = \nu$ and $\mathrm{Var}(X) = 2\nu$.

5 Suppose that observations $y_1, y_2, \ldots, y_n$ are generated from the autoregressive (AR) model of order one:

$$y_t = \alpha y_{t-1} + \epsilon_t,$$

where α is the AR parameter and ϵ_t is a Gaussian white noise with variance σ^2.

(i) Write down the likelihood function $L(\alpha, \sigma^2; y_{1:n})$ and the log-likelihood function $\ell(\alpha, \sigma^2; y_{1:n})$ of the parameters α and σ^2, based on observation $y_{1:n} = \{y_1, y_2, \ldots, y_n\}$. ***(3 marks)***

(ii) Using the approximation $\log(1 + x) \approx x$, for some x, with $|x| < 1$, show that the log-likelihood of part (i) can be approximated as

$$\ell(\alpha, \sigma^2; y_{1:n}) \approx -\frac{n}{2}\log(2\pi\sigma^2) - \frac{y_1^2}{2\sigma^2} + \left(\frac{y_1^2 - \sigma^2}{2\sigma^2}\right)\alpha^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \alpha y_{t-1})^2.$$

(7 marks)

(iii) Using (ii) and adopting unconditional least squares, show that the approximate likelihood estimates of α and σ^2 satisfy

$$\hat{\alpha} = \frac{\sum_{t=2}^{n} y_t y_{t-1}}{\sum_{t=3}^{n} y_{t-1}^2 + \hat{\sigma}^2} \quad \text{and} \quad \hat{\sigma}^2 = \frac{(1 - \hat{\alpha}^2)y_1^2 + \sum_{t=2}^{n}(y_t - \hat{\alpha} y_{t-1})^2}{n}.$$

(8 marks)

(iv) If $\sigma^2 = y_1^2$, show that the maximum likelihood of α based on unconditional least squares is approximately equal to the maximum likelihood of α using conditional least squares. ***(2 marks)***

6 An airliner on a long-haul journey is set to fly at a pre-specified steady speed of v km/hr, but atmospheric conditions sometimes speed it up and sometimes slow it down. A simple model for its true position x_t along the flight path at time t hours supposes that

$$x_{t+1} = x_t + v + \eta_t,$$

where $\{\eta_t\}$ is a zero mean normal white noise sequence with variance σ_η^2. The position x_t can be observed only with error, the observed position y_t at time t being

$$y_t = x_t + \epsilon_t,$$

where ϵ_t is a zero mean normal white noise with variance σ_ϵ^2, uncorrelated with the other variables.

(i) Show that if $\beta_t = (x_t, v)^\top$, the system can be described by the equations

$$y_t = g^\top \beta_t + \epsilon_t, \tag{1}$$

$$\beta_t = F\beta_{t-1} + \zeta_t, \tag{2}$$

where g is a constant vector, F is a constant matrix and ζ_t is a random vector. Give the values of g and F and write down the mean and covariance matrix of ζ_t. What are equations (2) and (3) called in the context of state space linear modelling? ***(4 marks)***

(ii) Suppose that hourly observations of the plane's position are available up to time $t-1$ and it is believed that the true position x_{t-1} at time $t-1$ has posterior normal distribution with mean m_{t-1} and variance v_{t-1}. Show that, given this information, the prior mean of the position x_t at time t is $m_{t-1}+v$ and that the prior variance is $v_{t-1}+\sigma_\eta^2$. ***(4 marks)***

(iii) Find the posterior distribution of the true position x_t at time t when an observation y_t of the plane's position at time t becomes available. ***(8 marks)***

(iv) Suppose that $v = 720$, $\sigma_\epsilon = 10$ and $\sigma_\eta = 30$. If, one hour into the flight, an observation gives the position as $y_1 = 750$ km and it was known with certainty that at time $t = 0$ the true position had been $x_0 = 0$, find the posterior mean of the plane's position at $t = 1$, and the associated posterior variance. Hence give an approximate 95% credible interval for the true position x_1 after one hour. ***(4 marks)***

End of Question Paper