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 2015-2016 **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. A population based five year study of live births has measured a number of risk factors for the outcome infant death (i.e. death before 1 year of age). A total of 259790 births were included and infant death occurred in 3557. The study collected data on a number of variables including maternal smoking (during pregnancy) and birthweight. Birthweight was categorised into low (less than 3kg), normal (3-3.5kg) and high (more than 3.5kg). The study results are summarised in the Table below. The study team is confident that birthweight is not on the causal pathway.

Birthweight	Mothers who smoked during pregnancy	Infant deaths for mothers who smoked	Mothers who did not smoke in pregnancy	Infant deaths for mothers who did not smoke	Risk of infant death when mother smoked in pregnancy	Risk of infant death when mother did not smoke in pregnancy	Relative Risk of infant death by maternal smoking (95% CI)
Under 3.0kg	8561	256	10855	434	0.030	0.040	0.748 0.643 – 0.871
3.0kg to 3.5 kg	27784	417	119384	1791	0.015	0.015	1.000 0.900 – 1.112
Over 3.5 Kg	6411	51	86821	608	0.008	0.007	1.133 0.853 – 1.507
Total	42770	724	217060	2833			

i) Define what is meant in epidemiology by 'a confounding variable' when using the counterfactual and the collapsibility definitions.

(4 marks)

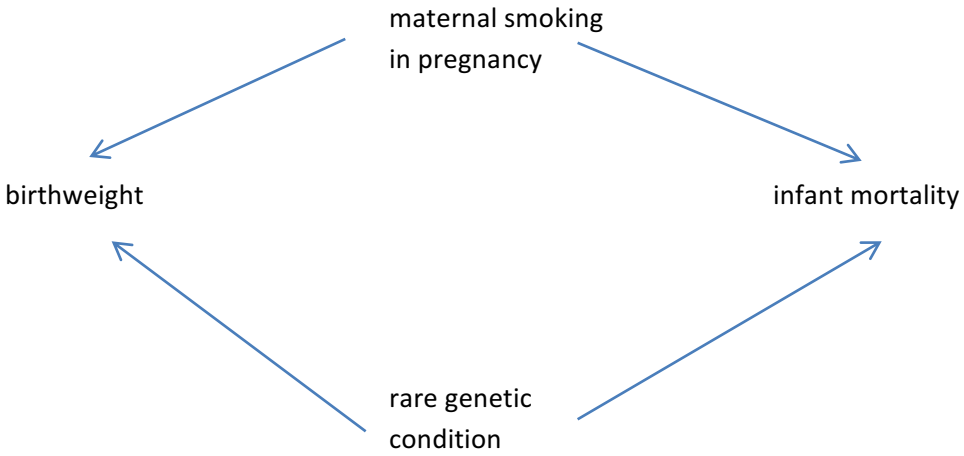
ii) The table above shows the relative risks of an infant death by maternal smoking stratified by birthweight of the baby. Calculate the crude relative risk of infant death by maternal smoking and report the 95% confidence interval.

(4 marks)

iii) It is assumed that the variable birthweight is **not** on the causal pathway between smoking and infant death. Use the counterfactual approach to calculate the Crude Expected Count (CEC) and the Standardised Expected Count (SEC) to assess the evidence that birthweight is a confounding variable. Report the Standardised Relative Risk (SRR).

(6 marks)

iv) There is evidence (external to this study) to suggest that a rare genetic condition results in lower birthweights and increases the risk of infant mortality. A causal graph summarising what is believed to be true about these causal relationships is shown in the Figure below.



In the light of the causal relations shown in the figure discuss whether it is reasonable to regard birthweight as a confounding variable.

(3 marks)

v) Taking into account all of the results and the causal graph above, comment on whether the stratified by birthweight (shown in the table), standardised over birthweights (SEC) or the crude relative risk is the most appropriate to report for a summary causal interpretation between maternal smoking in pregnancy and infant mortality.

(3 marks)

2. A 2011 study investigated the number of deaths caused by diseases known to be related to alcohol consumption. The aim of the study was to identify any differences in the rates of deaths between European countries. The table below shows the age-sex distribution of deaths caused by diseases known to be related to alcohol consumption in England in 2011. It also shows the corresponding age-sex distributions for population of England and the European Standard population.

Sex	Age Group	England		European Standard Population ('000s)
		Deaths	Population ('000,000s)	
Male	15-34	156	7.10	20.0
Male	35-54	1,849	7.30	14.0
Male	55-74	2,106	5.24	11.5
Male	75+	392	1.68	4.5
Female	15-34	100	7.00	20.0
Female	35-54	911	7.42	14.0
Female	55-74	977	5.51	11.5
Female	75+	284	2.46	4.5

i) Briefly explain why direct standardisation would be preferred to indirect standardisation when comparing mortality rates between countries.

(2 marks)

ii) Calculate the crude death rate per 100,000 for England.

(2 marks)

iii) Calculate the age and sex directly standardised rate per 100,000 for England.

(6 marks)

iv) Comment on the results from (ii) and (iii). What do these results tell you about the population of England and the European Standard Population?

(2 marks)

v) Calculate the standard error and 95% confidence interval for the directly standardised rate from (ii)

(6 marks)

vi) The directly standardised rate for one of the other countries in the study was found to be 16.2 per 100,000 with 95% confidence interval 15.8 to 16.5. Compare the rate from this country with the rate found for England. Is there evidence to suggest a difference between the two countries?

(2 marks)

3. Dapsone is a drug commonly used to treat infections and inflammations. Serious adverse reaction occurs in around 2% of patients treated with the drug. Two genome-wide association studies of patients who have suffered adverse reactions to Dapsone have found statistically significant associations with risk of an adverse reaction and the HLA-B13 allele. Dapsone is frequently used in the treatment of leprosy where a higher rate of adverse reactions has been reported.

A Chinese study has examined a cohort of 141 leprosy patients who were all treated with Dapsone and an ethnically similar control (without leprosy) population for association with the HLA-B13 allele. 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.

		Number of HLA-B13 alleles				
		0	**1**	**2**	**Row Total**	**HWE p-value**
Leprosy patients	**With adverse reaction**	6	30	2	38	0.0002
	No adverse reaction	88	14	1	103	0.603
Ethnically similar controls		794	151	6	951	0.621

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?

(3 marks)

iii) Compare the allele frequency of the HLA-B13 allele in patients with leprosy with an adverse reaction to the frequency of allele in those leprosy patients without an adverse reaction. Is there any evidence that HLA-B13 increases the risk of an adverse reaction to dapsone in leprosy patients?

(3 marks)

iv) Report appropriate comparative risk statistics for the comparative risk of an adverse outcome by each genotype compared with the baseline genotype of 'zero alleles'. Compare these two statistics (without calculating a 95% confidence interval) and comment whether the HLA-B13 allele appears to act in a dominant or codominant manner.

(4 marks)

v) Use an appropriate statistical test (assuming a two sided 5% significance threshold) to evaluate the evidence that the HLA-B13 allele is associated with risk of leprosy.

(4 marks)

vi) Discuss whether or not it is possible, using reasonable assumptions, to calculate the population attributable risk of suffering an adverse reaction to Dapsone due to the HLA-B13 risk allele from the data provided in this Chinese study.

(4 marks)

4 (i) A model is to be fitted to a time series of length 81. Values of the sample autocorrelation function (ACF) and sample partial ACF (PACF) are tabulated below.

Lag (h)	1	2	3	4	5
ACF (r_h)	$\star$	0.7	0.05	0.02	0.01
PACF ($\hat{a}_h^{(h)}$)	0.9	$\star\star$	0.4	-0.15	0.10

(a) Find the omitted values ($\star$ and $\star\star$). ***(3 marks)***

(b) Check whether the time series is stationary. ***(1 mark)***

(c) Test whether the time series is consistent with white noise. ***(2 marks)***

(d) Test whether the time series is consistent with MA(1) and MA(2) moving average models. ***(5 marks)***

(e) Test whether the time series is consistent with AR(1), AR(2), AR(3) and AR(4) autoregressive models. ***(3 marks)***

(f) Giving your reason, state which of the models in (d) and (e) you prefer for these data. ***(2 marks)***

(ii) Consider the time series model

$$y_t = 2\sin\left(\frac{\pi}{t}\right) + \epsilon_t, \quad t = 1, 2, \ldots,$$

where ϵ_t follows a white noise process with variance 10.

(a) Show that y_t is non-stationary process. ***(2 marks)***

(b) Define an appropriate transformation of y_t to result in a stationary time series model. Justify your choice. ***(2 marks)***

5 Consider that y_t is generated by an ARMA(1,1) model

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

where α, β are the AR and MA coefficients and ϵ_t is a Gaussian white noise with variance σ^2.

(i) Write down the likelihood and the log-likelihood functions of the parameters α, β and σ^2, based on a collection of observations $y_{1:n} = (y_1, y_2, \ldots, y_n)$.

(6 marks)

(ii) Using conditional least squares,

(a) derive the partial derivatives of the conditional log-likelihood with respect to α and β; ***(6 marks)***

(b) using part (a) show that the maximum likelihood estimates of α and β are

$$\hat{\alpha} = \frac{\sum_{t=2}^{n} y_t y_{t-1} \sum_{t=2}^{n} \epsilon_{t-1}^2 - \sum_{t=2}^{n} y_t \epsilon_{t-1} \sum_{t=2}^{n} y_{t-1} \epsilon_{t-1}}{\sum_{t=2}^{n} y_{t-1}^2 \sum_{t=2}^{n} \epsilon_{t-1}^2 - (\sum_{t=2}^{n} y_{t-1} \epsilon_{t-1})^2}$$

$$\hat{\beta} = \frac{\sum_{t=2}^{n} y_{t-1}^2 \sum_{t=2}^{n} y_t \epsilon_{t-1} - \sum_{t=2}^{n} y_{t-1} \epsilon_{t-1} \sum_{t=2}^{n} y_t y_{t-1}}{\sum_{t=2}^{n} y_{t-1}^2 \sum_{t=2}^{n} \epsilon_{t-1}^2 - (\sum_{t=2}^{n} y_{t-1} \epsilon_{t-1})^2}.$$

(8 marks)

6 Consider the ARMA(1,1) model for the time series y_t:

$$y_t = 0.2y_{t-1} + \epsilon_t + \epsilon_{t-1}, \tag{1}$$

where ϵ_t follows Gaussian white noise with variance $\sigma^2 = 1$.

Define the state vector

$$\beta_t = \begin{bmatrix} y_t \\ \epsilon_{t+1} \\ \epsilon_t \end{bmatrix}.$$

Using this state vector write down model (1) in state space form, i.e.

$$y_t = x\beta_t + \eta_t$$
$$\beta_t = F\beta_{t-1} + \zeta_t$$

and determine x, F, η_t, ζ_t and the variances of η_t and ζ_t. ***(3 marks)***

(i) If the posterior distribution of the state β_1, given $y_1 = 2$ is

$$\beta_1 \mid \{y_1 = 2\} \sim N\left\{ \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\},$$

then find the one-step forecast distribution of y_2. ***(6 marks)***

(ii) Using the result in (i) obtain a 95% predictive interval for y_2. ***(2 marks)***

(iii) At time $t = 2$ the observation y_2 is observed to be 3. Perform the Kalman filter iteration for $t = 2$ and obtain the posterior distribution of

$$\beta_2 \mid \{y_2 = 3\}.$$

(9 marks)

End of Question Paper