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 2014-2015 **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. Prior research into potential genetic risk factors for bladder cancer has found evidence of an association with a specific variant in the AKRO gene on chromosome 13q. The association was consistent with a dominant effect of the minor allele at the SNP rs22178. A new research team has investigated this SNP in a series of 1000 bladder cancer cases and 1000 cancer free controls sampled from a female cohort study. Bladder cancer is a smoking related cancer and the researchers want to examine the relation between bladder cancer, smoking and the risk SNP. The genotypes for the SNP rs22178 in cases and controls stratified by smoking status are shown in the table below.

	Number with genotypes AA/ AT/ TT	
	Cases	Controls
Ever Smokers	180/ 398/ 87	135/ 226/ 50
Never Smokers	208/ 104 /23	359/ 184/ 46
Total	388/ 502/ 110	494/ 410/ 96

i) Examine the quality of the genotyping in this study using the Hardy Weinberg test of equilibrium. Comment on your results.

(2 marks)

ii) Assuming the risk locus acts in a dominant manner and the minor allele is dominant to the major allele, calculate the appropriate comparative risk statistic with its 95% confidence interval and assess if this study confirms the original reported association between bladder cancer and rs22178.

(4 marks)

iii) Calculate the population attributable risk for this association between bladder cancer and the SNP rs22178 and interpret it. State any assumption you needed to make to perform this calculation.

(2 marks)

iv) Explore the relationship between the risk SNP (assuming any effects are dominant), smoking and bladder cancer and discuss whether smoking should be classed as a confounder, a modifier or neither of these. Use the 95% confidence intervals for each comparative statistic to weigh the evidence in your discussion.

(10 marks)

v) The study has also collected the ages of the participants. Suggest two methods or approaches to account for age as a possible confounding variable, when examining the relationship between smoking, rs22178 and bladder cancer in this study. Discuss the merits and disadvantages of each in light of your conclusions found in iv).

(2 marks)

2. A new biomarker based score has been proposed to pick up patients with early cancer. Twenty five biomarkers have been combined as they are individually sensitive to specific cancers. The score consists of counting how many biomarkers are positive in each tested subject. The final score belongs to one of four ordinal categories, <3, 3-8, 9-14 and 15+. The score test has been measured in a prospective sample of 1005 patients referred to a single cancer treatment hospital via their General Practitioner for investigation for presence of cancer. Following consent, each subject was measured using the biomarker score and then at the end of full medical investigations it was found that 305 of the 1005 did indeed have a cancer present. Those where no cancer was found initially were followed up for 36 months. If a cancer was diagnosed within this period they were classified as a case (there were 40 subjects who developed cancer during the follow-up period). The results of using this biomarker score on the 1005 patients are shown in the table below:

Biomarker score	Number with cancer	Number without cancer diagnosed
1 – 2	1	197
3 - 8	36	270
9 - 14	138	120
15+	170	73
Total	345	660

i) Calculate the sensitivity, specificity and diagnostic accuracy of this test for each of the three possible cut-points, i.e. greater than 2 , greater than 8 and greater than 14.

(6 marks)

ii) Assuming the optimal cut-off for the test is the one that gives the highest diagnostic accuracy, calculate the negative predictive value (NPV), likelihood ratio negative (LR-), positive predictive value (PPV) and the likelihood ratio positive (LR+).

(4 marks)

iii) Assuming the test (cut-off) with the highest diagnostic accuracy is to be used in practice, use the results from i) and ii) to discuss if this would be a useful test for screening (ruling in) and/or diagnosis (ruling out).

(4 marks)

iv) The researchers on the biomarker score are ultimately aiming to develop a test that could be used in the general population to assist in identifying those that need to be referred for further investigation. The table below shows the UK incidence of cancer per 100000 per year, stratified by age group. In the UK, a test will not be considered for national screening evaluation if there is less than a 10% chance that a person who is positive for the test has a true cancer present.
Use the information in the table below and the test characteristics calculated in i) at each of the three potential cut-offs to answer the following question. Do any of the cut-offs result in the score test meeting this UK screening criteria and if so which age groups in the population would the test be suitable for? Justify your answer.

(6 marks)

Age group	40-49	50-54	55-59	60-64	65-69	70-79	75-79	80-84
Average incidence per year per 100,000	289	465	701	1069	1518	1875	2307	2612

3. Researchers at the University of Sheffield recently conducted a study looking at hospital admissions for epilepsy related incidents in England. The aim of the study was to investigate if there are any differences in the rates of admissions between Primary Care Trusts. The table below shows the number of admissions and the age-sex population structure for one of the PCTs in the study along with the age-sex distribution of a standard population.

Age Group	Sex	PCT		Standard Population ('000,000s)
		Admissions	Population ('000s)	
18-34	Male	104	72	6.1
35-64	Male	252	84	10.3
65+	Male	99	29	3.8
18-34	Female	67	69	6.0
35-64	Female	133	86	10.5
65+	Female	75	37	4.8

i) Using the data provided in the Table, calculate the age and sex directly standardised rate per 100,000 for the PCT.

(5 marks)

ii) The observed number of admissions in the standard population for 2007 was 52,974. Calculate the Comparative Incidence Figure (CIF) for the PCT.

(1 mark)

iii) Using the standard error for the log-transformed CIF, calculate a 95% confidence interval.

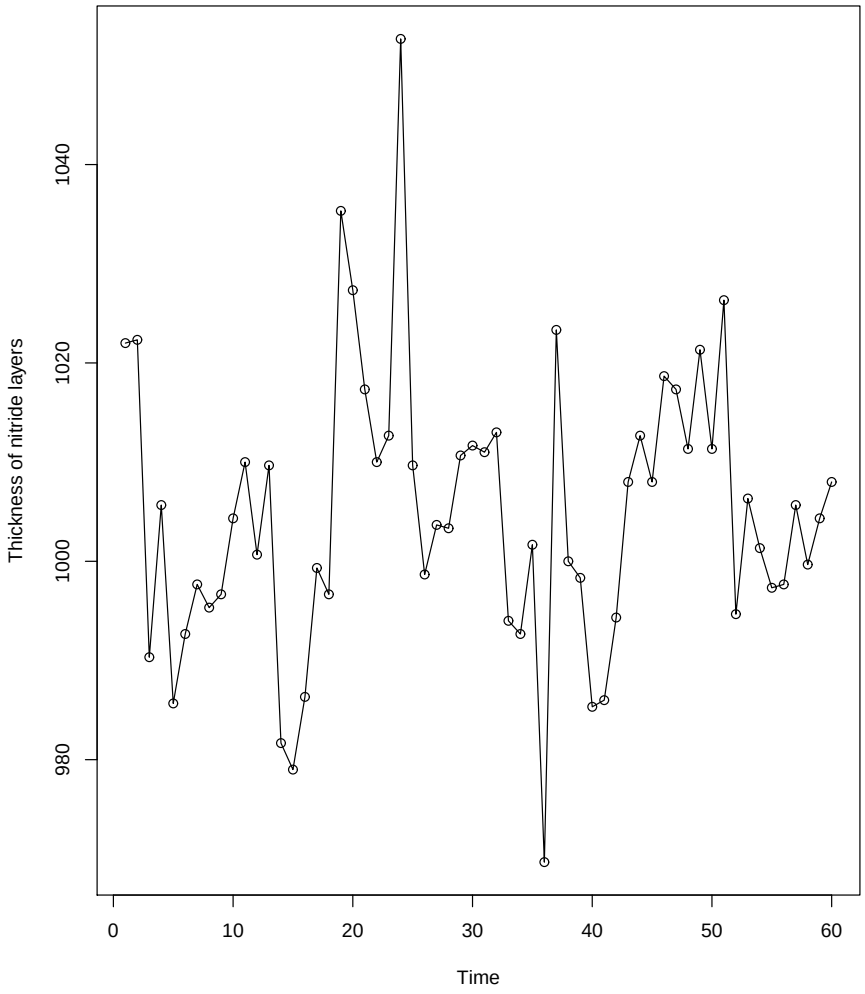
(6 marks)

iv) Comment on the results from (ii) and (iii). Is there evidence to suggest that the incidence rate in the PCT is different to that of the standard population? Please justify your answer.

(4 marks)

v) The research team also requested similar data from a second PCT but found that the number of emergency admissions for females aged 65+ was missing. What would be the impact of this missing data when calculating the directly standardised incidence rate? If these missing data were ignored what impact would it have when comparing the rates in the two PCTs? Suggest how a standardised rate may be calculated allowing for missing data.

(4 marks)



4 (i) The plot above shows time series data consisting of 60 observations of the thickness of nitride layers (unknown units); the data was part of a larger experiment on the manufacturing of a microelectronic device [1].

(a) Describe the data by commenting on their structure, their variation and dynamics. ***(2 marks)***

(b) Based on your answer in (a) or otherwise, suggest suitable time series model(s) that may be appropriate for this data. ***(2 marks)***

[1]Source: Triantafyllopoulos, K., Godolphin, J.D. and Godolphin, E.J., 2005, Process improvement in the microelectronic industry by state space modelling, Quality and Reliability Engineering International, 21, 465-475

4 (continued)

(ii) A model is to be fitted to a time series of length 100. 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)	0.6	0.4	0.1	0.05	0.01
PACF ($\hat{a}_h^{(h)}$)	$\star$	$\star\star$	0.02	0.01	-0.02

(a) Find the omitted values ($\star$ and $\star\star$). ***(4 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 moving average models. ***(4 marks)***

(e) Test whether the time series is consistent with autoregressive models. ***(3 marks)***

(f) Based on your answer in (c)-(e) above, suggest a time series model that may be suitable to model the data. ***(2 marks)***

5 (i) In the context of maximum likelihood estimation of ARMA models describe briefly what is meant by conditional least squares estimation. ***(2 marks)***

(ii) Consider that y_t is generated by an autoregressive model of order 2 (AR(2))

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t,$$

where α_1, α_2 are the AR coefficients and ϵ_t is white noise with variance σ^2.

(a) Write down the likelihood and the log-likelihood functions of the parameters α_1, α_2 and σ^2, based on a collection of observations $y_{1:n} = (y_1, y_2, \ldots, y_n)$. ***(4 marks)***

(b) Using conditional least squares, show that the maximum likelihood estimates of α_1, α_2 and σ^2 are

$$\hat{\alpha}_1 = \frac{\sum_{t=3}^n y_{t-2}^2 \sum_{t=3}^n y_t y_{t-1} - \sum_{t=3}^n y_{t-1} y_{t-2} \sum_{t=3}^n y_t y_{t-2}}{\sum_{t=3}^n y_{t-1}^2 \sum_{t=3}^n y_{t-2}^2 - \left(\sum_{t=3}^n y_{t-1} y_{t-2}\right)^2}$$

$$\hat{\alpha}_2 = \frac{\sum_{t=3}^n y_{t-1}^2 \sum_{t=3}^n y_t y_{t-2} - \sum_{t=3}^n y_t y_{t-1} \sum_{t=3}^n y_{t-1} y_{t-2}}{\sum_{t=3}^n y_{t-1}^2 \sum_{t=3}^n y_{t-2}^2 - \left(\sum_{t=3}^n y_{t-1} y_{t-2}\right)^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{t=3}^n \left(y_t - \hat{\alpha}_1 y_{t-1} - \hat{\alpha}_2 y_{t-2}\right)^2 .$$

(14 marks)

6 A company trades 10 products, with the ith product projected to give a return r_{it} at time t, for $i = 1, 2, \ldots, 10$. The company believes that each of these returns r_{it} follows an autoregressive process

$$r_{it} = 0.9 r_{i,t-1} + \zeta_{it},$$

where ζ_{it} is a white noise with variance 1, $\zeta_{it} \sim N(0, 1)$, and ζ_{it} is independent of ζ_{jt}, for any $i \neq j$.

Due to a data recording error r_{it} is not available. However, the aggregate return can be observed subject to additive noise, according to the model

$$y_t = \sum_{i=1}^{10} r_{it} + \epsilon_t,$$

where ϵ_t is a white noise with variance 1, $\epsilon_t \sim N(0, 1)$, and it is assumed that ϵ_t is independent of ζ_{it}, for any t and for any i.

(i) Define the state

$$\beta_t = \sum_{i=1}^{10} r_{it}.$$

Show that y_t follows a state space model

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

and determine x, F, ζ_t and the variance of ζ_t. ***(4 marks)***

(ii) A prior distribution for β_0 is set as

$$\beta_0 \sim N(0, 100).$$

If the first observation is $y_1 = 2$, perform the Kalman filter iteration for $t = 1$ and obtain the posterior distribution of

$$\beta_1 \mid \{y_1 = 2\}.$$

(8 marks)

(iii) Using the result in (ii) obtain a 95% predictive interval for y_2. ***(4 marks)***

(iv) Describe briefly what is the likely effect on the posterior distribution of β_t (for large t), if the prior distribution of β_0 changes from (a) $\beta_0 \sim N(0, 1)$ to (b) $\beta_0 \sim N(0, 1000)$. ***(4 marks)***

End of Question Paper