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 2011-2012 — **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. It is believed that a certain gene increases the risk of skin damage when a person is exposed to the sun. A population survey by a dermatologist examined whether people had evidence of permanent skin damage due to the sun and asked then about serious sun burn in the past. A saliva sample was taken for genetic testing. The results were as follows:

	Gene positive		**Gene negative**	
	Sun burn	No sun burn	Sun burn	No sun burn
Skin damage	45	30	1200	300
No skin damage	5	20	1800	2700
	50	50	3000	3000

(i) According to the classic definition of confounding, is there evidence of confounding between the presence and absence of the gene and being sun-burnt?

(2 marks)

(ii) Derive the table that would have been obtained if we had had no data from the genetic testing.

(4 marks)

(iii) Find the risk difference, relative risk and odds ratio of sun burn for skin damage for the situations where the gene is positive, the gene is negative and for the collapsed table.

(6 marks)

(iv) Comment on whether there is evidence of confounding according to the collapsibility definition of confounding.

(2 marks)

(v) Find the standardised expected count of those expected to develop the disease in the sunburnt cohort as if they had not been sunburnt. Compare with the crude expected count of those not sunburnt and show that in fact there is no confounding.

(6 marks)

2. Researchers at Sheffield Children's hospital recently published a paper looking at the effect of collagen gene polymorphisms on fracture risk and bone mass acquisition during childhood (Blades et al, Bone, 2011). They recruited 378 children, of whom 195 had sustained a fracture. As part of the study they measured the genotypes at the COL1A2 SNP in order to determine whether these were related to fracture risk. Examining the data for the COL1A2 SNP, of the 195 children with a fracture, 18 were homozygous for the minor allele P, and 83 were homozygous for the major allele p. Of the 183 children without a fracture 34 were homozygous PP and 77 were homozygous pp.

(i) What type of study is this? Please justify your answer

(3 marks)

(ii) Display these data in a suitable table. Are the genotype frequencies observed consistent with this SNP being in Hardy Weinberg Equilibrium at the population level?

(3 marks)

(iii) Previous studies have suggested that the risk of fracture is similar when individuals carry one or two copies of the major allele p. Hence the data can be reduced to a 2 x 2 table for analysis. Discuss the relative merits and/or disadvantages of using an allele based test and/or the Cochran Armitage trend test in the light of this prior knowledge.

(3 marks)

(iv) Calculate a suitable measure of the risk of fracture for the children who are homozygous PP compared to those who carry at least 1 p allele together with its 95% confidence interval. Comment on the results.

(6 marks)

(v) The researchers were concerned that a second SNP COL1A could be a confounder. Discuss what is meant by a confounder in this context.

(2 marks)

(vi) The odds ratio for fracture for those children who were homozygous for COL1A minor allele S was 0.28 (0.11 to 0.75) and the odds ratio for those who carried at least one s allele 0.65 (0.29 to 1.46). Explain the implications of this finding for the analysis of the COL2 gene and state how you would deal with any problem that might arise. ***(3 marks)***

3. A genome wide linkage study has been performed on 66 siblings pairs recruited in the UK, where both siblings are affected with autism. The genome wide study generated genotypes at 316 highly polymorphic markers. This genotyping was then used to calculate the identical by descent (IBD) sharing between the sibling pairs. The strongest signal for linkage was found at the marker D7S522, located on chromosome 7q.

(i) At the marker D7S522 the IBD sharing distribution was as follows:
5% of sibling pairs shared no alleles IBD, 50% shared exactly one allele IBD and 45% shared two alleles IBD. What is the statistical evidence of linkage to this locus? Is this statistically significant at the genome wide linkage significance threshold?

(4 marks)

(ii) This chromosomal region has been further studied in a family based association study. The chromosomal region contains the MET gene and 35 SNP polymorphisms were genotyped within this gene in an international collection of 494 parent offspring trios. Each trio consisted of a single affected child with autism and both of the child's parents. The strongest evidence of non random transmission of alleles to the offspring was with the SNP rs38841. A summary of the genotyping observed in all 494 trios for this single SNP is presented in Figure 1. Use the transmission disequilibrium test (TDT), to formally evaluate the evidence for association with this SNP and risk of autism. Identify, with justification which allele is potentially associated with a higher risk of autism. Is your conclusion altered when you consider this is only one of 35 considered SNPs?

(6 marks)

iii) A neighbouring SNP to the rs38841 is in complete linkage disequilibrium (ie D'=1). This SNP has been genotyped in the same trios as above but the p-value resulting from the TDT is larger than for rs38841. Suggest two reasons for this result.

(2 marks)

(iv) The MET gene has been studied in other populations and there is some suggestion that there is different viability or fitness associated with different genotypes within MET. One study found that genotype specific relative viabilities for GG (relative to AA) was 1.05 and AG (relative to AA) was 1.1. A sample of newborns were measured for this SNP and the following frequencies were observed:

AA: 34, AG: 77, GG: 21.

(a) Calculate the allele frequency of the G allele in the sample. Using this as your best estimate of the assumed to be large population value, what allele frequency is predicted by the system after two generations?

(4 marks)

(b) Reparameterise this relative viability system into a model using the 'selection coefficient' and 'heterozygous effect'. Comment on the predicted long term behaviour and presence of this polymorphism in future generations.

(4 marks)

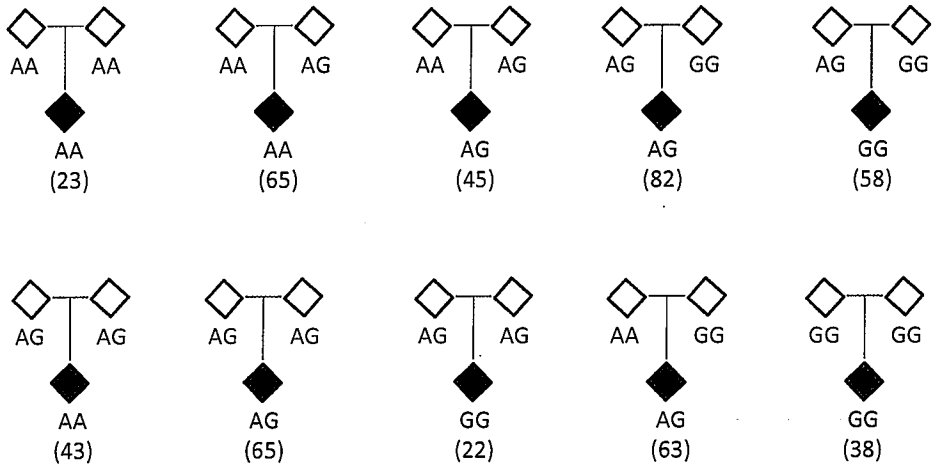

Figure 1: Numbers of trios observed with specified genotype configuration for rs38841.
Gender of parents and affected offspring is not identified.
Integer in () indicates number of trios observed with this genotype configuration.

4 (i) The figure below shows the quarterly averages of the Euro/£ and US$/£ exchange rates between 2001 and 2005 (source: Bank of England). Also plotted is the price of gold relative to its price in the first quarter of 2001. Describe the three time series and their relationship, using suitable technical terms and adding approximate quantification where appropriate. Detailed numerical comparisons of the series are not required. *(4 marks)*

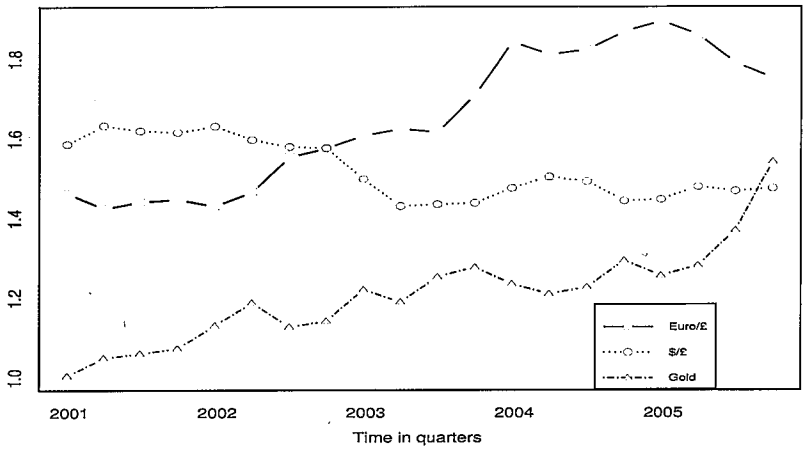

Figure 1: Dollar/Pound, Euro/Pound and Gold Index, Jan 2001 - Dec 2005

(ii) The following table shows the sample acf r_h and pacf $a_h^{(h)}$ values of the Dollar/pound series where the lag $h = 1, 2, 3, \ldots$ is in quarters of a year.

Lag h	1	2	3	4	5	6	7	8
r_h	0.9123	*	0.6287	0.4604	0.2703	0.0898	-0.0653	-0.2012
$a_h^{(h)}$	*	-0.3586	-0.0071	-0.2779	-0.1745	-0.0563	-0.0252	-0.0706

Calculate the two missing values. Without making any further calculations, describe briefly how using the above table, you could investigate whether the time series is stationary or not. *(7 marks)*

(iii) Investigate and compare possible models for the series given the values in the table in (ii) and your calculations for the missing ones. What further evidence is desirable for a more complete identification of a suitable model? *(9 marks)*

5 Consider the time series model

$$X_t = \frac{1}{3}X_{t-1} + \epsilon_t + \frac{1}{2}\epsilon_{t-1} - \frac{1}{4}\epsilon_{t-2},$$

where ϵ_t is white noise with variance 4.

(i) Explain why in this model X_t has zero mean. *(2 marks)*

(ii) Show that this model is invertible. *(3 marks)*

(iii) If $Var(X_t) = 9$, calculate the autocorrelation function (ACF) of X_t. *(10 marks)*

(iv) Find a state space representation for the model for X_t. Write down the observation and evolution equations and state the distribution of the observation and evolution innovations. *(5 marks)*

6 (i) Consider the AR(1) time series model

$$X_t = \alpha X_{t-1} + \epsilon_t,$$

where ϵ_t is a white noise sequence with variance σ^2.

(a) Derive the formula

$$X_{n+k} = \alpha^k X_n + \sum_{i=0}^{k-1} \alpha^i \epsilon_{n-i}$$

for some positive integers n and k. *(2 marks)*

(b) Using (a) or otherwise, derive the k-step ahead forecast mean and variance of X_{n+k}, based on observed time series data $X_1 = x_1, \ldots, X_n = x_n$. *(7 marks)*

(c) If ϵ_t is normally distributed, write down a 95% prediction interval for X_{n+k}. *(1 mark)*

(ii) Consider the AR(2) time series model

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \epsilon_t,$$

where ϵ_t is as above a white noise sequence with variance σ^2.

(a) Write down X_t in state-space form. *(1 mark)*

(b) Using the state-space form in (a), derive the 2-step ahead forecast mean and variance of X_{n+2}, based on observed time series data $X_1 = x_1, \ldots, X_n = x_n$. *(7 marks)*

(c) Hence show that the 2-step forecast variance of X_{n+2} is the same as that of an AR(1) model defined by $X_t = \alpha_1 X_{t-1} + \epsilon_t$. *(2 marks)*

End of Question Paper