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

Academic Session 2010-11

Epidemiology and Time Series

3 Hours

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. In the paper by Talmud et al (2010, BMJ) the Cambridge risk score (CRS) was calculated to predict the risk of developing Type 2 diabetes in middle aged people over the following 10 years.
Suppose a random sample of 10 people who developed diabetes and 10 who did not develop diabetes gave the following results:

CRS
Developed Diabetes: 0.01, 0.12, 0.15, 0.18, 0.24, 0.26, 0.27, 0.30, 0.31, 0.32
Did not develop Diabetes: 0.02, 0.03,0.05, 0.07, 0.08, 0.09, 0.10, 0.19, 0.20, 0.28

a) Find the odds ratio of being above a risk of 0.275 in the two groups. Can we use this to estimate the risk of developing diabetes given a CRS over 0.275?

(2 marks)

b) Find the sensitivity of the test for a specificity of 90%

(2 marks)

c) Plot the ROC using the three quartiles of the distribution of CRS as cut-off points

(4 marks)

d) Find the area under the ROC curve found in (iv) and interpret it.

(4 marks)

e) What statistics would be useful to an individual to decide if they were at risk from Diabetes and can you derive them from this table. If not, why not?

(4 marks)

f) What other factors have to be in place for the CRS to be a useful screening tool?

(4 marks)

2) School children in Bavaria have their height and weight measured routinely at the time of entry to school. In 1997 researchers used a subsample of these data to examine whether there was a link between breast feeding and obesity. They found that the prevalence of overweight children in the 4,022 children who were never breastfed was 12.6%, whilst the prevalence of overweight children in the 5,184 children who were breastfed at some point was 9.2%. An overweight child was defined as one having a BMI above the 90th centile. (R von Kries et al. BMJ 1999; 319:147-150)

a) What type of study is this?

(1 mark)

b) Which of the following two would be most suitable measure to summarise the outcome of this study: relative risk or odds ratio?

(1 mark)

c) Using your answer from (b) calculate one of these statistics and its confidence interval. Note that you are comparing the 'risk' of being overweight for breast fed children compared to children who were not breastfed. Is this value statistically significantly different from the null value? Please justify your answer.

(7 marks)

The researchers looked at the duration of breast-feeding and found the following results:

Table 1: Prevalence of overweight (95% confidence intervals) by duration of breastfeeding

	Prevalence of overweight children (%) (95% CI)
Never breastfed	12.6(12.4 to 12.9)
Exclusively breast-fed:	
<= 2 months	11.1(10.6 to 11.6)
3-5 months	8.4 (8.1 to 8.8)
6-12 months	6.8 (6.1 to 7.6)
>12 months	5.0 (1.1 to 8.8)

i) Does table 1 suggest that there might be a relationship between the duration of breast-feeding and the likelihood of being overweight? Can you suggest how you might examine this more formally?

(3 marks)

ii) What do you notice about the confidence intervals in the table above. Can you think of an explanation for this pattern?

(2 marks)

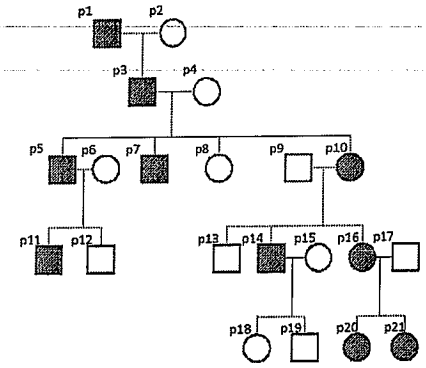
iii) Based on the evidence thus presented do you think that breastfeeding has a protective effect on the risk of being overweight in childhood? Please justify your answer

(6 marks)

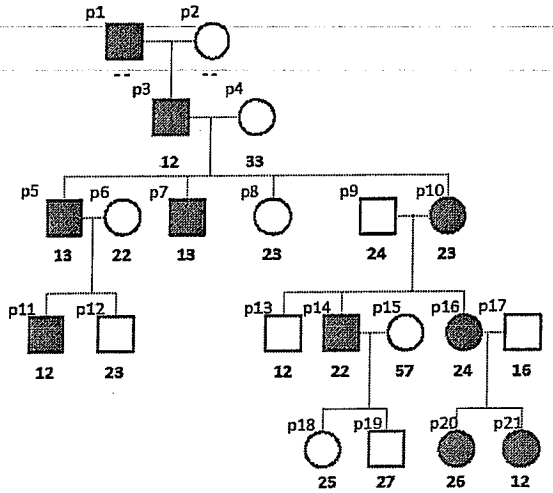
3a) The family represented in Figure 1a contains individuals diagnosed with a rare fully penetrant dominant disorder. Those affected with the disorder are shaded, unshaded squares or circles indicate the person is unaffected. Each person is uniquely identified through a label shown at the top left of each symbol.

Figure 1 (a) Family tree of individuals diagnosed with a rare fully penetrant dominant disorder and (b) with genetic marker

a)



b)



i) Identify and list the founders in this pedigree.

[1 mark]

Consider the general segregation of alleles through the pedigree at a distinct autosomal locus which has not been genotyped or observed in this family.

ii) What is the probability that p11 and p16 share 0, 1 or 2 alleles identical by descent?

[2 marks]

iii) if p12 and p18 were to mate and produce an offspring, what is the probability that the offspring would receive two alleles identical by descent?

[2 marks]

Figure 1b presents the same family as Figure 1a but now genotyped for the genetic marker (named D3S311 and with seven distinct observed alleles) located on chromosome 3p. The genotype is shown written below each individual in the family, where '- -' indicates that DNA was not available from that person so the genotype could not be directly observed. This genetic marker and the pedigree will be used to evaluate the evidence for linkage between the underlying dominant disease locus and this locus on chromosome 3p.

iv) List the individuals who are homozygous for the marker D3S311.

[1 mark]

Question 3 continued on next page

v) How many informative meioses does the family genotyped in Figure 2 present? How many recombinations have occurred between the two loci considered in this problem. Justify your answer.

[2 marks]

vi) Stating your assumptions derive the LOD score function, in terms of the recombination fraction θ, for the pedigree and linkage information.

[2 marks]

vii) The maximum likelihood estimate of the recombination fraction for this case is 0.09. What is the statistical evidence for linkage to this region?

[2 marks]

3b)

A candidate SNP (named rs546789) has been investigated for association with risk of acute lymphoblastic leukemia (ALL). An incident series of 245 cases (all diagnosed under age 16) were recruited from a major cancer hospital for genetic study. Parents and unrelated hospital controls (a total of 248) were also recruited. The genotyping results of the candidate SNP (with alleles A and G) are summarised in Table 2 below.

Table 2: Family trio genotyping for rs546789:

Father	Mother	Affected child	Total families of this type
AA	AA	AA	4
AA	AG	AA	9
AA	AG	AG	4
AA	GG	AG	12
AG	AA	AA	12
AG	AA	AG	5
AG	AG	AA	19
AG	AG	AG	10
AG	AG	GG	13
AG	GG	AG	28
AG	GG	GG	21
GG	AA	AG	11
GG	AG	AG	29
GG	AG	GG	18
GG	GG	GG	50

Hospital Control series genotyping for rs546789:

Genotype	AA	AG	GG
Count:	38	69	141

Question 3 continued on next page

i) Association studies can be conducted by recruiting family units or unrelated individuals from the population. Contrast the benefits and disadvantages of these two approaches.

[2 marks]

ii) Evaluate the evidence for an association between the risk of ALL and variation at the SNP rs546789, by performing a family based transmission disequilibrium test and a case-control allelic distribution test. Comment on your results and how the data is used differently by the two tests.

[4 marks]

iii) Examine the hospital control sample for consistency with Hardy Weinberg Equilibrium (HWE). Comment on your results to part b) in the light of the HWE analysis.

[2 marks]

Question Four Continues on Next Page

4 (i) (a) In the context of descriptive analysis of time series x_t, briefly explain why a moving average for even span s is *not* defined as

$$\frac{1}{s}(x_{t-s/2} + x_{t-s/2-1} + \cdots + x_{t-1} + x_t + x_{t+1} + \cdots + x_{t+s/2-1}).$$

(1 mark)

(b) Consider the time series with values

$$x_1 = 5, \quad x_2 = 4, \quad x_3 = 6, \quad x_4 = 5, \quad x_5 = 7, \quad x_6 = 6, \quad x_7 = 3.$$

Using the *correct* definition of the even-span moving average, calculate moving averages of span 4, for the values x_3, x_4 and x_5. *(3 marks)*

(ii) A time series of length 70 gave values for the sample autocorrelation function (ACF), denoted by r_h and values for the partial ACF, denoted by a_h, according to the table below.

Lag h	1	2	3	4
r_h	0.58	0.43	0.37	0.22
a_h	*	*	0.19	0.21

(a) Using this table, find the values of a_1 and a_2, indicated in the table by stars. *(4 marks)*

(b) Test whether this time series is consistent with a white noise process, a moving average model and an autoregressive model. *(10 marks)*

(c) Suggest a model which you would expect to fit well to this time series data. *(2 marks)*

5 Consider the time series model

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

where ϵ_t is a white noise process with variance 3, i.e. $\epsilon_t \sim WN(0, 3)$.

(i) Give the abbreviated name of the model for X_t. *(1 mark)*

(ii) Write down model (1) in compact form, using the backward shift operator B. *(2 marks)*

(iii) Show that model (1) is causal and invertible. *(5 marks)*

(iv) Find the variance of X_t. *(12 marks)*

6 Consider the trend dynamic linear model, given by equations

$$X_t = [1,0]\begin{bmatrix}\theta_{1t}\\ \theta_{2t}\end{bmatrix} + \epsilon_t = \boldsymbol{F}^T\boldsymbol{\theta}_t + \epsilon_t, \tag{2}$$

$$\boldsymbol{\theta}_t = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t = \boldsymbol{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \tag{3}$$

where $\boldsymbol{\theta}_t = [\theta_{1t}, \theta_{2t}]^T$ is a state vector, ϵ_t follows a normal distribution with zero mean and variance 50, and $\boldsymbol{\omega}_t$ follows a bivariate normal distribution with zero mean vector and covariance matrix

$$\boldsymbol{W} = \begin{bmatrix}20 & 0\\ 0 & 20\end{bmatrix},$$

written as

$$\boldsymbol{\omega}_t \sim N_2\left\{\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}20 & 0\\ 0 & 20\end{bmatrix}\right\}.$$

It is also assumed that ϵ_t and $\boldsymbol{\omega}_t$ are mutually and individually independent, and they are independent of the initial state $\boldsymbol{\theta}_0$. Suppose that $x_1, x_2, \ldots, x_n$ values of the time series are observed and that the posterior distribution of $\boldsymbol{\theta}_n$, given information $x^n = (x_1, \ldots, x_n)$ is given by

$$\boldsymbol{\theta}_n | x^n \sim N_2\left\{\begin{bmatrix}250\\ 100\end{bmatrix}, \begin{bmatrix}10 & 0\\ 0 & 33\end{bmatrix}\right\}.$$

For some positive integer $k > 0$, define the new time series

$$S_n = X_{n+1} + X_{n+2} + \cdots + X_{n+k}.$$

(i) Show that the k-step forecast function of $\{X_t\}$ is $\widehat{X}_{n+k} = E(X_{n+k}|x^n) = 100k + 250$. *(4 marks)*

(ii) Find the posterior mean of S_n, given x^n, for $k = 2$. *(2 marks)*

(iii) For $k = 2$, show that, given x^n, the covariance of X_{n+1} and X_{n+2} is 96, and hence calculate the posterior variance of S_n, given x^n. *(13 marks)*

(iv) Derive the posterior distribution of S_n, given x^n, for $k = 2$. *(1 mark)*

End of Question Paper