RESTRICTED OPEN BOOK EXAMINATION
Data provided: Table of χ^2 quantiles

MAS371

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2014–2015**

Applied Probability **2 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
Marks will be awarded for your best **three** *answers. Total marks 60.*

1 (i) The general Markov chain with two states, labelled 0 and 1, has transition matrix

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

where p and q are unknown parameters. A sequence of observations of the state of the chain $X_0, X_1, X_2, \ldots, X_N$ is made.

(a) Write down the log-likelihood $l(P|X_0, \ldots, X_N)$ in terms of summaries

$$N_{ij} = \#\text{ observed transitions from state } i \text{ to state } j$$

and

$$N_{i\cdot} = \sum_{j=0,1} N_{ij}$$

and hence find the observed information $J(P|X_0, \ldots, X_N)$.

(4 marks)

(b) Show that $(q/(p+q), p/(p+q))$ is a stationary distribution for P.

(1 mark)

(c) Show that if X_0 has the distribution in (b), then

$$\mathrm{E}[N_{11}] = Nq(1-q)/(p+q).$$

(3 marks)

Question 1 continued on next page

1 (continued)

(ii) The symmetric case of the model in part (i) has transition matrix

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

(a) Write down the log-likelihood and observed information for p based on $X_0, X_1, X_2, \ldots, X_N$. ***(2 marks)***

(b) Show that in this case the expected information can be calculated without any assumption about the distribution of X_0. ***(3 marks)***

(iii) Now consider the case where the state of the symetric chain in part (ii) is observed only after every two time steps, so the data are $X_0, X_2, X_4, \ldots, X_N$ (assuming N is even, for simplicity).

Show that the likelihood for p is of the form

$$(2p(1-p))^m \left((p^2 + (1-p)^2\right)^n$$

in terms of suitable summaries m, n of the data, which you should define. ***(3 marks)***

Hence show that $p = 1/2$ is always a solution of the likelihood equation. Explain how you could check whether this solution represents the maximum likelihood estimate in a given case. ***(2 marks)***

Without carrying out this calculation, explain whether or not you think $p = 1/2$ is the maximum likelihood estimate

(a) if $m \ll n$

(b) if $m \geq n$. ***(2 marks)***

2 As part of a study of the welfare of farm animals, the behaviour of two cows was monitored throughout the course of several hours. Each cow's behaviour was classified every 10 minutes into Feeding, Resting or Other. The tables below show the numbers of pairs of consecutive observations, categorised by behaviour, for cow A and cow B separately.

Cow A		Behaviour after		
		Feeding	Resting	Other
Behaviour before	Feeding	4	3	1
	Resting	4	16	3
	Other	1	3	1

Cow B		Behaviour after		
		Feeding	Resting	Other
Behaviour before	Feeding	1	2	2
	Resting	3	11	7
	Other	1	7	2

(i) Assuming that cow A's transitions between behaviours at 10 minute intervals can be modelled as a Markov chain, give maximum likelihood estimates and estimated standard errors for the elements of its transition matrix P_A. ***(5 marks)***

Calculate the estimated standard error for the difference between the probability that an observation of Resting is followed by Resting again and the probability that it is followed by Feeding. ***(3 marks)***

(ii) Comment briefly on the similarities and differences in the tables of counts for the two cows. ***(2 marks)***

Carry out a generalized likelihood ratio test of the hypothesis that both cows' behaviours follow independent Markov chains with the *same* transition matrix, say P_C, against the more general alternative that they follow independent Markov chains of the same form but with different transition matrices P_A and P_B respectively. ***(8 marks)***

You may wish to use the fact that, writing n_{ij} and m_{ij} for the observed numbers of transitions for cows A and B respectively, as tabulated above, and $n_{i\cdot}, m_{i\cdot}$ for the corresponding row sums, we have

$$\sum_{ij} m_{ij} \log(m_{ij}/m_{i\cdot}) = -33.93,$$

$$\sum_{ij} (n_{ij} + m_{ij}) \log((n_{ij} + m_{ij})/(n_{i\cdot} + m_{i\cdot})) = -67.74,$$

(iii) It is suspected that the two cows are not behaving independently of each other. Explain briefly how a Markov chain model could be constructed that allowed investigation of this dependence. ***(2 marks)***

3 The reproduction of a certain species of micro-organism can be modelled as a pure birth process $X(t)$ in which the birth rate is $\lambda x(x-1)$ when the population size is x.

(i) Write down an expression for the probability

$$P(X(t+\delta t) = x+1 | X(t) = x)$$

(2 marks)

(ii) If the population is fully observed from time 0 to time t, write down the likelihood for the parameter λ in terms of the observed times spent at each population size $a_x, x = 0, 1, \dots$ and the total number of births observed, b. Derive expressions for the maximum likelihood estimate of λ and the observed information $J(\lambda)$. (You may assume standard results for a general continuous-time Markov chain.) ***(4 marks)***

(iii) In an automated laboratory experiment, the changes over time in the population are summarized in terms of the initial population $X(0)$, the duration of the experiment t, the final population $X(t)$, the time average of the population size, defined as

$$A = \sum_x \{(a_x/t)x\},$$

and the variance over time of the population, which can be defined as

$$V = \sum_x \{(a_x/t)x^2\} - A^2.$$

(a) Show that the likelihood from part (ii) can be written in terms of these summaries. ***(4 marks)***

(b) If one experiment gives $t = 60$ (in minutes), $X(0) = 20$, $X(t) = 100$, $A = 50$, $V = 400$, estimate λ and give an approximate 95% confidence interval for it. ***(6 marks)***

(c) Give a rough estimate of how much longer the experiment would need to continue in order to give a relative standard error (e.s.e.$(\hat{\lambda})/\hat{\lambda}$) of less than 0.1. ***(4 marks)***

4 A sample of lake sediment contains grains of pollen from a rare tree species. Let v_i be the date of the ith grain, $i = 1, \ldots, n$, measured from the oldest part of the sediment, and let t be the known latest date within the sediment.

(i) A simple model for the distribution over time of the pollen grains is a homogeneous Poisson process of rate λ. If the sample covers 2000 years, so $t = 2000$, and $n = 400$, state the maximum likelihood estimate of λ and its estimated standard error. ***(2 marks)***

Explain briefly why the values of $v_1, \ldots v_n$ are not needed for these calculations. ***(1 mark)***

(ii) A more realistic model incorporates the fact that the tree is believed to have become extinct in the region, at some time τ during the interval represented by the sample, with $0 \leq \tau \leq t$. This leads to an inhomogeneous Poisson model, with

$$\lambda(u) = \begin{cases} \alpha & 0 \leq u \leq \tau \\ 0 & \tau < u \leq t. \end{cases}$$

Write down the likelihood $L(\alpha, \tau | v_1, \ldots, v_n)$ for this model, taking care to specify the region of the parameter space over which it is non-zero. (Hint: it may help to consider the summary statistic $w = \max\{v_1, \ldots, v_n\}$.) ***(4 marks)***

Give an expression for the log-likelihood over the region where the likelihood is positive. ***(1 mark)***

If $w = 1500$, and t and n are as before, find maximum likelihood estimates of α and τ, and an estimated standard error for α only. ***(5 marks)***

(iii) Since n is large, α in the inhomogeneous model can be estimated quite well. Treating α as if it were fixed at $\hat{\alpha}$, calculate a likelihood interval for τ of the form

$$\{\tau : l(\tau) - l(\hat{\tau}) > -c\}$$

with $c = 2$, and comment briefly. ***(5 marks)***

(iv) If it is believed that the local extinction may have been a gradual process, rather than a sudden event, explain how the model could be modified to incorporate this. ***(2 marks)***

End of Question Paper

Table of the pth quantile of the χ^2 distribution with ν degrees of freedom, $\chi^2_{p,\nu}$

		ν								
		1	2	3	4	5	6	7	8	9
	0.10	0.016	0.211	0.584	1.064	1.610	2.204	2.833	3.490	4.168
	0.50	0.455	1.386	2.366	3.357	4.351	5.348	6.346	7.344	8.343
p	0.90	2.706	4.605	6.251	7.779	9.236	10.645	12.017	13.362	14.684
	0.95	3.841	5.991	7.815	9.488	11.070	12.592	14.067	15.507	16.919
	0.99	6.635	9.210	11.345	13.277	15.086	16.812	18.475	20.090	21.666