UNIVERSITY OF KENT

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

LEVEL H EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

Monday, 14 May 2007: 2.00 – 4.00

This paper is divided into TWO *sections as follows:*

Section A*: Six short questions each marked out of 10. Candidates may attempt all* SIX *questions but are advised that they cannot obtain more than* FORTY *marks, in total, on this section.*

Section B*: Four longer questions each marked out of 30. Candidates may not attempt more than* TWO *of the* FOUR *questions in this section.*

Candidates are advised to show their working on their scripts. Marks might then be allocated for use of a correct method, even if the numerical or algebraic result is incorrect.

Copies of the New Cambridge Elementary Statistical Tables *are provided.*
Approved calculators may be used.

SECTION A

These questions will each be marked out of 10. Candidates may attempt all SIX *questions but are advised that they cannot obtain more than* FORTY MARKS*, in total, on this section.*

1. Let $x_1, ..., x_n$ denote a random sample from the Cauchy distribution, with single unknown parameter θ, which has probability density function given by

$$f(x) = \frac{1}{\pi\{1 + (x - \theta)^2\}}, \quad -\infty < x < \infty.$$

Show that the Newton-Raphson iterative method for obtaining the maximum likelihood estimate $\hat{\theta}$ has the form below, where $\hat{\theta}^{(m)}$ denotes the mth iterate for $\hat{\theta}$:

$$\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} - \frac{\sum_{i=1}^{n} \frac{(x_i - \hat{\theta}^{(m)})}{\{1+(x_i-\hat{\theta}^{(m)})^2\}}}{\sum_{i=1}^{n} \frac{(x_i-\hat{\theta}^{(m)})^2-1}{\{1+(x_i-\hat{\theta}^{(m)})^2\}^2}}.$$

[10 marks]

2. The data in the table below describe the incidence of menarche in groups of Polish girls.

Mean age of group (years)	9.21	11.58	13.08	15.33	17.58
No. having menstruated	0	10	47	107	1049
No. of girls	376	105	99	111	1049

Let μ denote the age by which 50% of Polish girls have reached menarche. Explain how you would use a logit model to obtain the maximum likelihood estimate of μ. (You do not need to calculate $\hat{\mu}$.) [10 marks]

3. Given below is a MATLAB function for calculating a kernel density estimate. Explain how the function works. The function calls a function *delta*. Explain the rôle played by *delta*, and suggest a possible form that it might take.

```
function z=kernel(y,data,k1)
n=length(data);
h=k1*std(data)/n^0.2;
z=0;
for i=1:n
 z=z+delta((y-data(i))/h);
end
z=z/(n*h);
```

[10 marks]

4. A religious sect in the seventeenth century performed educational plays, and printed programmes for these plays in batches of size a.

   (a) For a particular batch, if p denotes the probability that any programme survives to the present day, and it is assumed that individual programmes survive independently, write down the probability that i copies of a programme survive from that batch. [5 marks]

   (b) As plays differed in their popularity, we can approximately model the distribution of sizes of batches of play programmes by the Poisson distribution, with mean λ. Let X be the number of surviving programmes from a randomly selected play. Show that X has the Poisson distribution with parameter $\beta = \lambda p$. [5 marks]

5. The distribution of the sizes of surviving batches of theatre programmes is Poisson with parameter β. It is found that 711 batches survive with only 1 copy, 83 with 2, 5 with 3, 4 with 4, 1 with 5, and that no batches survive with more than 5 copies. Write down an appropriate likelihood, taking account of the fact that the number, N, of batches that have no copies surviving is not known. [10 marks]

6. Let p_0 denote the probability that a batch of theatre programmes has no surviving copies. If N denotes the total number of batches of theatre programmes printed, and $p_0 = e^{-\beta}$, then the following relationship holds: $804 = \hat{N}(1 - \hat{p}_0)$. For one set of data, the maximum likelihood estimate of β and the estimate of standard error, are given by $\hat{\beta} = 0.26(0.024)$. Obtain an estimate of N, and use the $\delta-$method to obtain an estimate of the variance for the resulting estimate of N. [10 marks]

SECTION B

These questions will each be marked out of 30. Candidates may not attempt more than TWO *of the* FOUR *questions.*

7. A random sample of data on the remission times of leukaemia patients result in m uncensored times, $\{x_i, i = 1, \ldots, m\}$, and n right censored times, $\{x_i, i = m+1, \ldots, m+n\}$.

(a) Explain in detail how the EM algorithm could be used to fit an exponential model, with probability density function $f(x) = \rho e^{-\rho x}$, to these data using the method of maximum likelihood. [12 marks]

(b) Write down an expression for the likelihood. Explain how the likelihood would be modified if it was known that the response time of the $(m+1)$th patient was $< \tau$. [11 marks]

(c) A particular sample is given below, where a * indicates that the value is right-censored and all times are measured in weeks.

6* 6 6 6 7 9* 10* 10 11* 13 16 17* 19* 20* 22 23 25* 32* 32* 34* 35*.

A graph of the log-likelihood function is shown below. Use this graph to obtain an approximate 95% confidence interval for ρ. [7 marks]

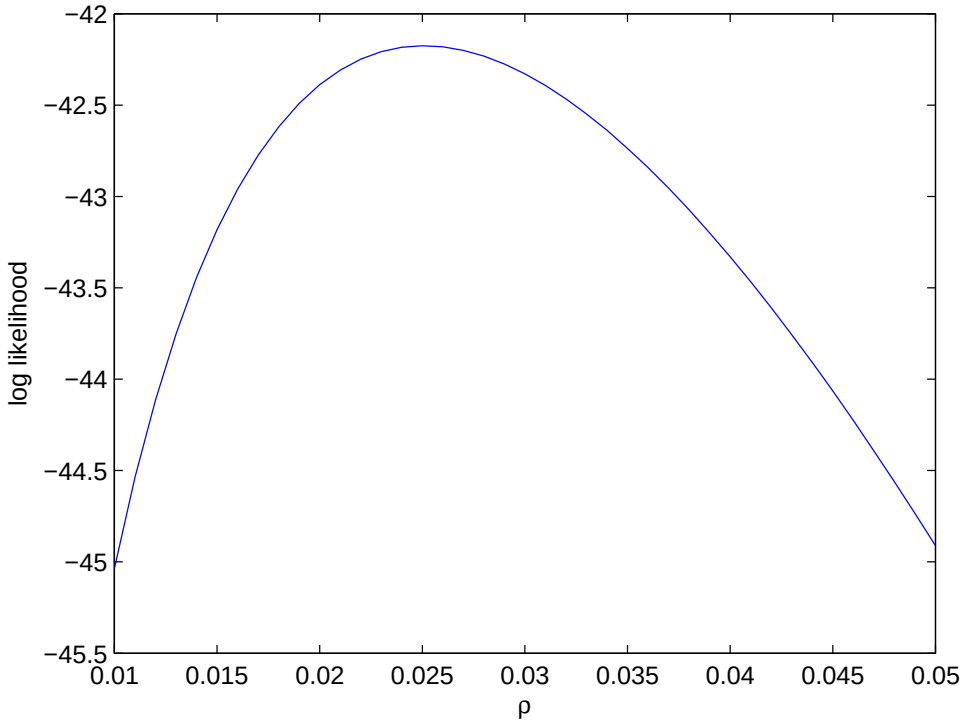

Figure 1: Log-likelihood for Question 7.

8. Times to conception are collected retrospectively on n couples that are successful in conceiving, resulting in the data, $\{n_i, i = 1, \ldots, r\}$, and n_{r+1}, so that n_i couples each take i cycles to conceive, and n_{r+1} couples take more than r cycles to conceive. A geometric distribution is proposed for these data, with probability function

$$p_i = p(1-p)^{i-1}, \qquad \text{for} \quad i \geq 1.$$

(a) Write down the likelihood and obtain the maximum likelihood estimate of p. If ℓ denotes the log-likelihood, use the fact that

$$\mathrm{E}\left[\frac{d^2\ell}{dp^2}\right] = -\frac{n\{1-(1-p)^r\}}{p^2(1-p)}$$

in order to obtain the asymptotic variance of the maximum-likelihood estimator, $\hat{p}$

[15 marks]

(b) A particular data set corresponding to $r = 12$ and $n = 100$ is given below.

Cycles	1	2	3	4	5	6	7	8	9	10	11	12	> 12
Frequency	29	16	17	4	3	9	4	5	1	1	1	3	7

When a modified Poisson distribution is fitted to these data, the expected frequencies are as follows:

3.91 12.67 20.54 22.20 17.99 11.67 6.30 2.92 1.18 0.43 0.14 0.04 0.01.

Perform a Pearson X^2 test of the goodness of fit of the modified Poisson distribution to the conception data. Discuss the relative merits of the two models for the data, and include in your answer consideration of more complex alternatives. [15 marks]

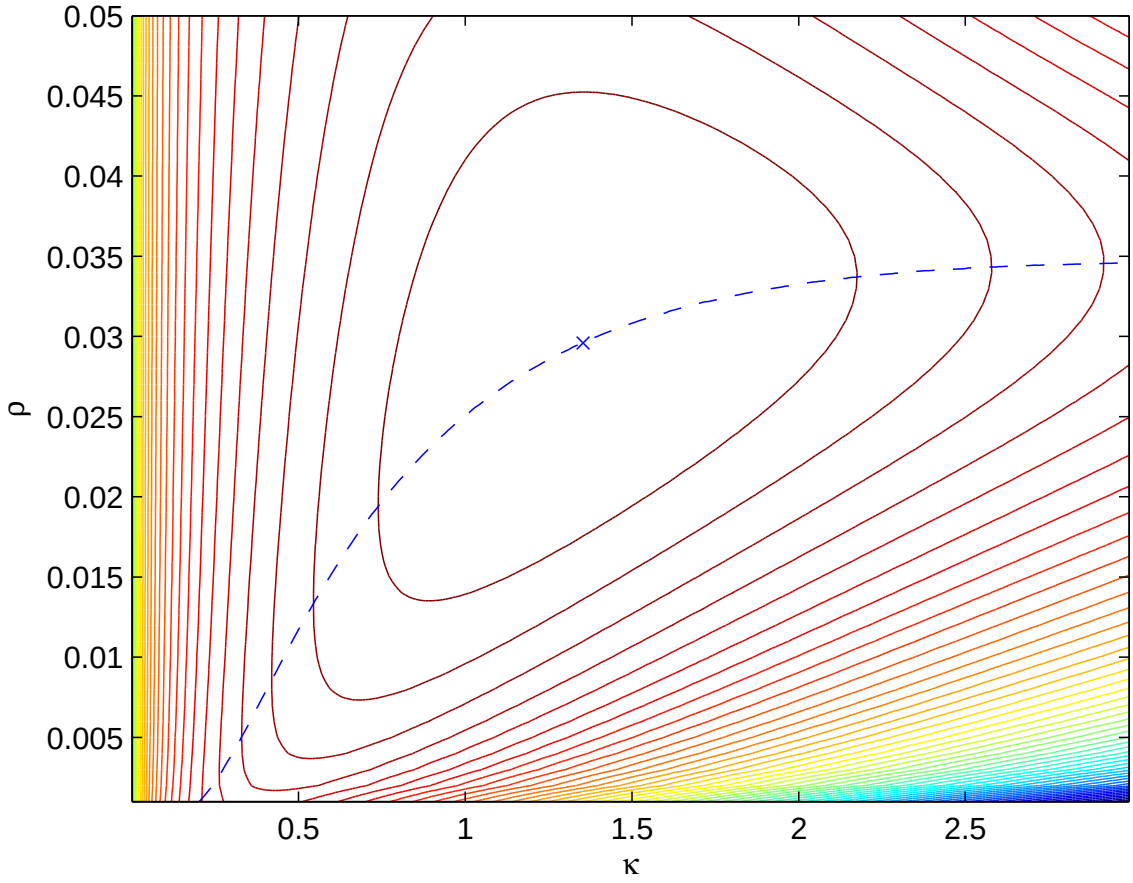

Figure 2: Log-likelihood for Question 9.

9. The Weibull distribution has the probability density function,

$$f(x) = \kappa\rho(\rho x)^{\kappa-1} exp\{-(\rho x)^{\kappa}\}, \quad \text{for} \quad x > 0.$$

(a) When the Weibull distribution is fitted to leukaemia remission time data, we obtain the log-likelihood surface illustrated in Figure 2.

Also shown in this figure is the location of the maximum-likelihood estimate of the two model parameters and the values of ρ corresponding to the profile log-likelihood with respect to κ. Explain how this profile log-likelihood curve is formed, provide a rough sketch of its shape, and describe how it may be used. [15 marks]

(b) It is desired to undertake a score test of the hypothesis that $\kappa = 1$, ie., that the exponential distribution is adequate to describe the leukaemia data. The score function under this hypothesis is given by

$$s = m + \sum_u \log x_i - m\frac{\sum x_i \log x_i}{\sum x_i}.$$

Here m is the number of uncensored terms in the sample and $\sum_u$ denotes summation for only the uncensored terms. Otherwise the summation is over all censored and uncensored times. Explain, in outline only, how the score function is derived. [6 marks]

Also under the exponential hypothesis, the appropriate leading term in the inverse of the observed information matrix has the form

$$v = (I_{\kappa\kappa} - I_{\kappa\rho}^2/I_{\rho\rho})^{-1}.$$

For the leukaemia data, we have $I_{\kappa\kappa} = 15.79$, $I_{\kappa\rho} = -246.0$, $I_{\rho\rho} = 14320.0$, $\sum_u \log x_i = 21.19$, $\sum x_i = 359$, and $\sum x_i \log x_i = 1077.3$.

Form the score test statistic, $s^2 v$, and decide whether or not the data support the added complexity of the Weibull distribution. [9 marks]

10. The MATLAB program opposite produces the three index plots shown in Figure 3.

(a) In each line after a for command a random variable Y is simulated. Derive the probability density function for this random variable, and explain which general method of simulation is being used. [12 marks]

(b) Each index plot is the result of the operation of a particular MCMC method. Explain in detail which method is being used. In addition, state the distribution of the random variable being simulated by the MCMC method. [12 marks]

(c) Discuss the different forms of the three index plots, explaining what they represent, and which you would use. [6 marks]

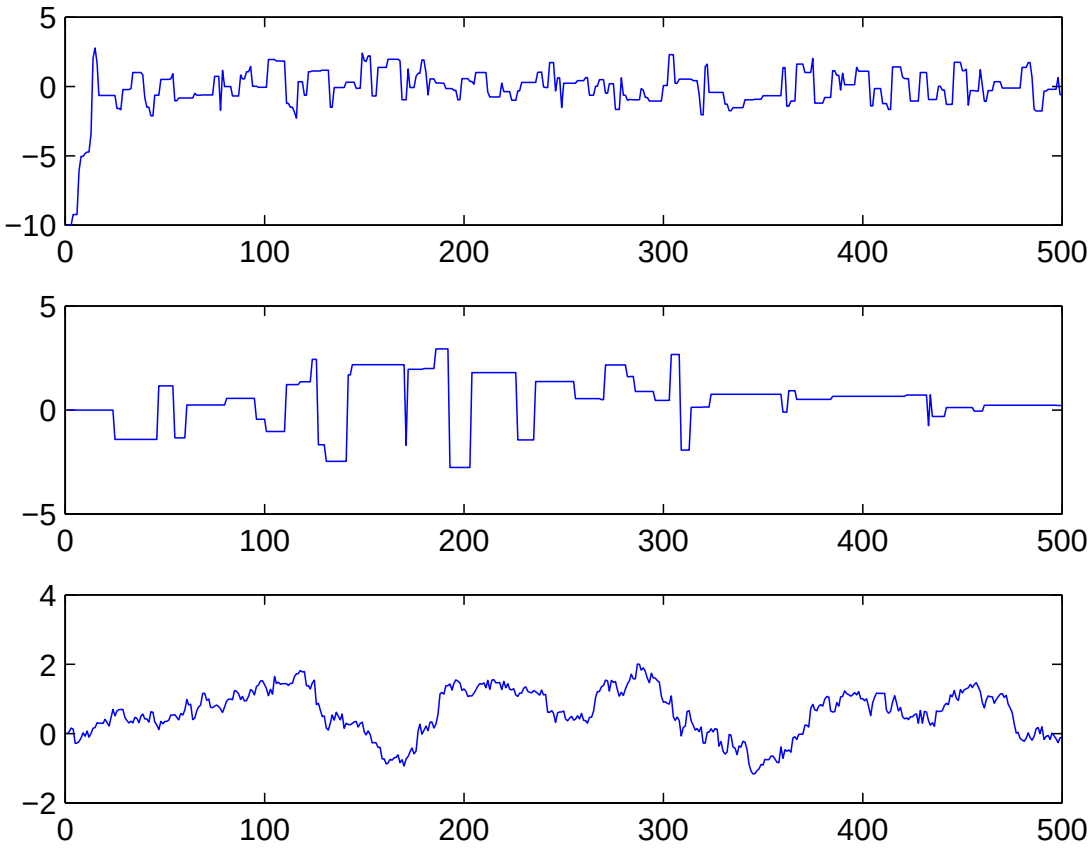

Figure 3: Index plots for Question 10.

```
sig1=0.5; sig2=0.1; sig3=10;n=500;
X1=zeros(1,n); X2=X1; X3=X1;
X1(1)=-10; X2(1)=0; X3(1)=0;
for i=2:n
y=-(1/sig1)*log(1/rand-1)+X1(i-1);
u=rand(1);
normpdf=@(x,mu,sigma)(exp(-0.5*(x-mu)^2/(sigma^2)))/(sqrt(2*pi)*sigma);
alpha=normpdf(y,0,1)/normpdf(X1(i-1),0,1);
if u <= alpha
    X1(i)=y;
else
    X1(i)=X1(i-1);
end
end
for i=2:n
y=-(1/sig2)*log(1/rand-1) +X2(i-1);
u=rand(1);
normpdf=@(x,mu,sigma)(exp(-0.5*(x-mu)^2/(sigma^2)))/(sqrt(2*pi)*sigma);
alpha=normpdf(y,0,1)/normpdf(X2(i-1),0,1);
if u <= alpha
    X2(i)=y;
else
    X2(i)=X2(i-1);
end
end
for i=2:n
y=-(1/sig3)*log(1/rand-1) +X3(i-1);
u=rand(1);
normpdf=@(x,mu,sigma)(exp(-0.5*(x-mu)^2/(sigma^2)))/(sqrt(2*pi)*sigma);
alpha=normpdf(y,0,1)/normpdf(X3(i-1),0,1);
if u <= alpha
    X3(i)=y;
else
    X3(i)=X3(i-1);
end
end
```