UNIVERSITY OF KENT

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

LEVEL H EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

22 May 2009: 9.30 a.m. – 11.30 a.m.

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.*

Copies of the New Cambridge 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. The data below give the number of fertility cycles to conception for 486 human couples.

198 107 55 38 18 22 7 9 5 3 6 6 12

Thus 198 couples took 1 cycle, 107 took 2, ... 6 took 12 cycles and 12 took at least 13.

(i) Explain how to allocate probabilities to each of the elements of the data vector under the geometric model. [3 marks]

(ii) Devise a way of estimating the probability, p, of conception per cycle by taking appropriate ratios of successive terms in the data vector. [3 marks]

(iii) Provide two further alternative ways of estimating p, and indicate which you regard as best overall, giving your reasons. Detailed calculation is not required. [4 marks]

2. Fitted to the data of Question 1 by the method of maximum likelihood, the beta-geometric model produces estimates of the model parameters, $\hat{\mu} = 0.408(0.020)$ and $\hat{\theta} = 0.137(0.032)$.

(i) Explain the terms in parentheses. [2 marks]

(ii) These terms are obtained in MATLAB by application of the following commands

```
x=fminsearch('betageo', x0);
h=hessian('betageo', x);
cov=inv(h);
k=sqrt(diag(cov));
```

Explain, in outline only, the operation of each of these commands. [5 marks]

(iii) What statistical theory is being invoked here? [3 marks]

3. The MATLAB program given below plots a log-likelihood which results when a Poisson Process of parameter λ is fitted to polyspermy data on the sea urchin, *Echinus esculentus*. Eggs are exposed to sperm in four separate experiments, and fertilisation is stopped at the times given in the vector `t`. These correspond in order to the rows of the matrix `m`, the entries of which describe the numbers of eggs with 0,1, ... fertilisations.

```
clear
m=[89 11 0 0 0 0;
    42 36 6 0 0 0;
    28 44 7 1 0 0;
    2 81 15 1 1 0];
t=[5 15 40 180];
lambda=0.005:0.0001:0.02;
logl=0;
for i=1:4
    for k=1:6
        p = exp(-lambda*t(i)).* (lambda*t(i)).^(k-1)/factorial(k-1);
        logl=logl+m(i,k)*log(p);
    end
end
plot(lambda, logl)
```

(i) Write down an algebraic expression for the likelihood. Explain how this expression relates to the formation of the log-likelihood in the above MATLAB program.

[5 marks]

(ii) Given in Figure 1 is the resulting log-likelihood graph. Explain how you would use it to obtain a 95% confidence interval for λ. Exact calculation is not required.

[5 marks]

4. Four stochastic models are proposed for the failure times of springs, which were tested for reliability. The resulting models are compared in the table below, in terms of maximised log-likelihood, number of model parameters (p) and two information criteria, AIC and BIC.

Model	p	-Max. log-likelihood	AIC	BIC
M_1	12	360.40	744.8	*
M_2	7	378.90	*	*
M_3	2	411.50	*	*
M_4	2	460.56	*	929.3

(i) Complete the table. [7 marks]

(ii) Deduce which model is the best for the data, giving your reasons. [3 marks]

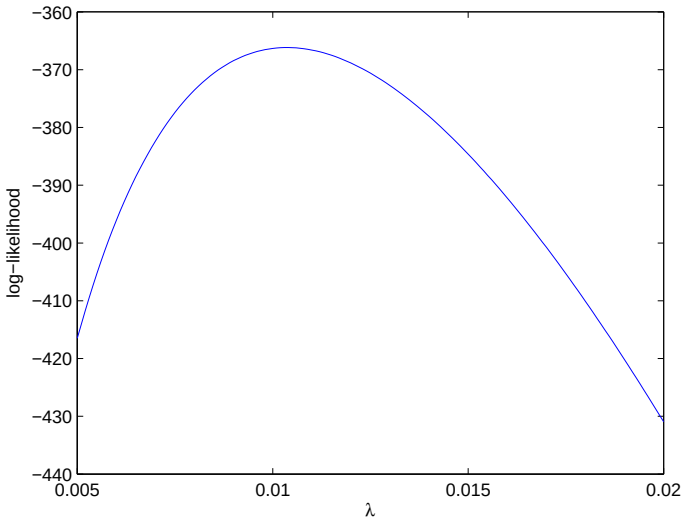

Figure 1: Figure for Question 3.

5. The MATLAB program below performs a deterministic numerical optimisation in order to produce maximum-likelihood estimates.

```
global data start gr
gr=1;
data=[1.09 -0.23 0.79 2.31 -0.81];
start = [.5 .5];
while norm(gr) > 0.001
  gr=grad('cauchy',start);
  lambda=fminbnd('linesearch', -0.5 0.5);
  start=start+lambda*gr';
end
```

(i) Explain which method is being used. [5 marks]

(ii) The function `linesearch` contains two lines, the first of which is the `global` statement to match that in the displayed program. Write down the second line. [5 marks]

6. The data below describe the mortality of adult flour beetles, *Tribolium confusum*, after 5 hours exposure to gaseous carbon disulphide.

Dose(mg/l)	49.06	52.99	56.91	60.84	64.76	68.69	72.61	76.54
No. of beetles	59	60	62	56	63	59	62	60
No. killed	6	13	18	28	52	53	61	59

(i) Explain how you would model these data using a logistic distribution. [3 marks]

(ii) Provide two different parameterisations of the model, one based on the ED_{50}, and explain the relationship between the two. [4 marks]

(iii) Explain how you would summarise the data in terms of the ED_{50}, including its estimated standard error. No detailed calculation is required. [3 marks]

SECTION B

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

7. The data below give the frequency distribution over household size of the total number of people infected in households of size 3 in a measles epidemic.

Total number infected	1	2	3
Frequency	34	25	275

A stochastic model predicts that the total number of people infected will have the following distribution,

Total number infected	1	2	3
Probability	$(1-\theta)^2$	$2\theta(1-\theta)^2$	$\theta^2(3-2\theta)$

where θ is an unknown probability of adequate contact.

(i) Obtain a quadratic equation that is satisfied by the maximum-likelihood estimate $\hat{\theta}$, and solve to show that $\hat{\theta} = 0.728$. [9 marks]

(ii) Carry out a Pearson goodness-of-fit test to examine if the model provides a good fit to the data. [8 marks]

(iii) A more refined model splits those epidemics in which 3 people are infected into two types, $3(a)$ and $3(b)$. The corresponding probability distribution is then shown below.

Final outcome	1	2	3(a)	3(b)
Probability	$(1-\theta)^2$	$2\theta(1-\theta)^2$	$2\theta^2(1-\theta)$	θ^2

For n independent observed epidemics, we obtain the frequency distribution below.

Final outcome	1	2	3(a)	3(b)
Frequency	a	b	c	d

If the values of c and d are not known, show how you would use the EM algorithm to maximise the likelihood. [9 marks]

(iv) Will the resulting estimate of θ still be 0.728? [4 marks]

8.

(i) Provide an outline graphical description of the Wald, score and likelihood-ratio tests, giving illustrations. How are the three tests related? [10 marks]

(ii) A different, smaller, data set from that of Question 1, but also giving the number of fertility cycles to human conception, with the same right-truncation at 12 cycles, is given below

29 16 17 4 3 9 4 5 1 1 1 3 7

When these data are fitted by a beta-geometric model, the maximum-likelihood estimates of the two model parameters are $\hat{\mu} = 0.2775(0.036)$ and $\hat{\theta} = 0.091(0.060)$. These values are evidently different from those given in Question 2, for the data in Question 1, but we need to test whether they are significantly different.

Making use of the maximised log-likelihood values given below, conduct a likelihood-ratio test of the null-hypothesis that the two data sets can be described by the same beta-geometric model. [10 marks]

Model	–Max log-likelihood
small data set	218.77
large data set	890.39
data sets combined	1115.80

(iii) Suppose that you want to test the same null-hypothesis as is tested above in (ii) by a likelihood-ratio test, but this time you prefer to use a score test. Explain how you would do this. [10 marks]

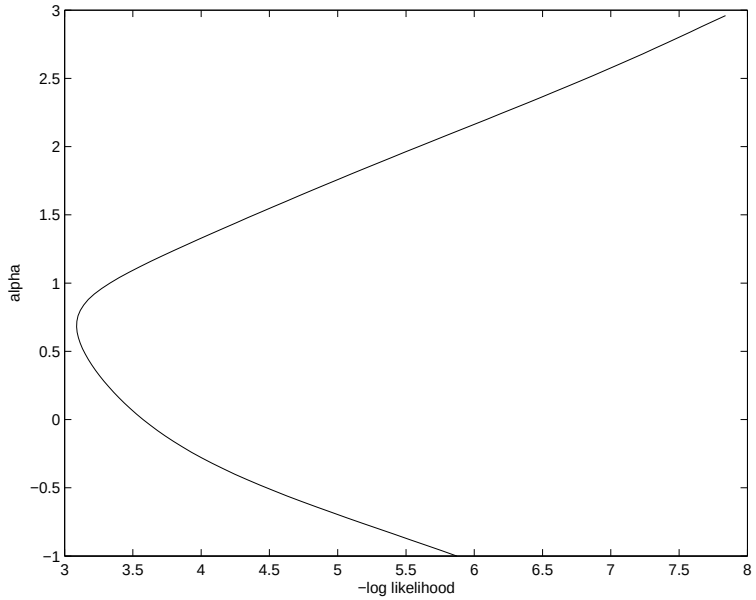

Figure 2: Profile log-likelihood for Question 9.

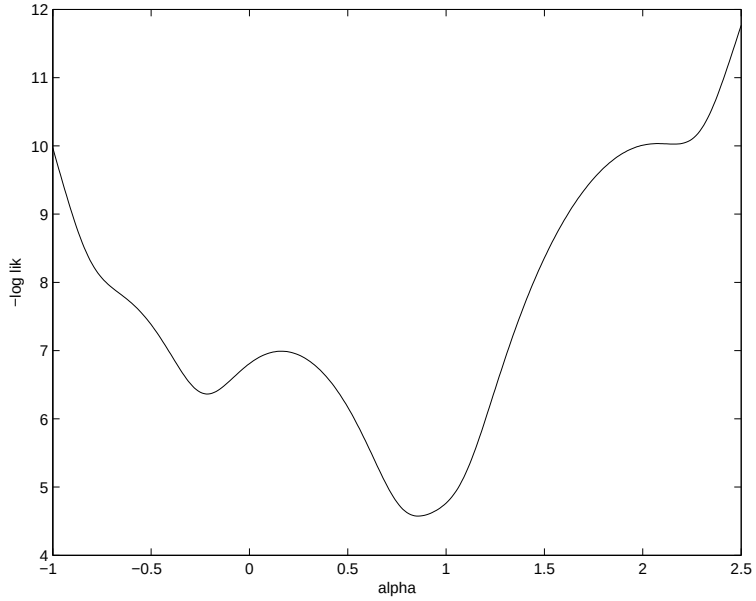

Figure 3: Section of log-likelihood surface, when $\beta = 0.2$, for Question 9.

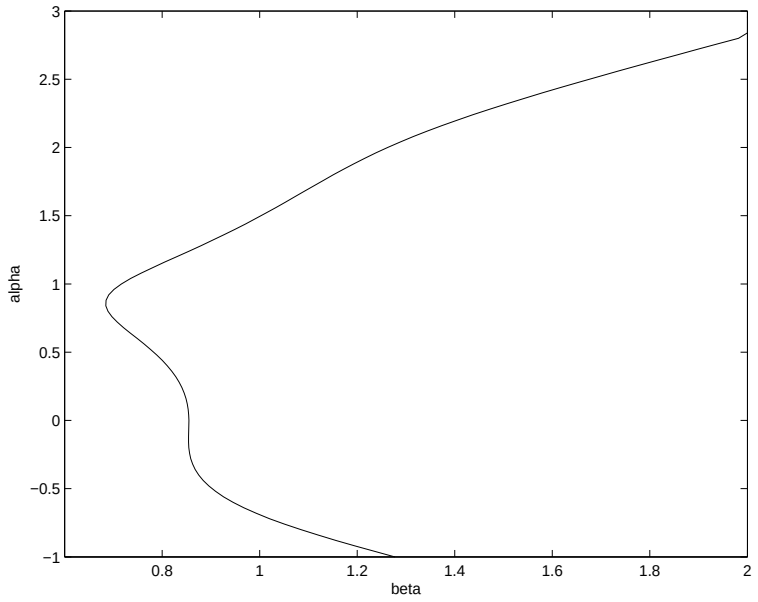

Figure 4: Path traversed in the parameter space corresponding to the profile log-likelihood, for Question 9.

9. The two-parameter Cauchy distribution has probability density function given by

$$f(x) = \frac{\beta}{\pi\{\beta^2 + (x-\alpha)^2\}}, \quad \text{for} \quad -\infty < x < \infty.$$

(i) Write down an expression for the log-likelihood corresponding to a random sample $\boldsymbol{x}$ of size n. [6 marks]

Corresponding to the particular case of $n = 5$ and $\boldsymbol{x} = \{1.09, -0.23, 0.79, 2.31, -0.81\}$, we obtain the graphs of Figures 1-3. The first of these graphs is a profile log-likelihood; the second is a section of the log-likelihood surface corresponding to $\beta = 0.2$, and the third is the path traversed in the parameter space corresponding to the profile log-likelihood.

(ii) Explain the difference between Figures 2 and 3, and explain the relevance of the path of Figure 4 to the profile of Figure 2. [10 marks]

(iii) Describe how you would obtain a confidence interval for α from the profile of Figure 2. [6 marks]

(iv) Explain the connection between the log-likelihood section of Figure 3 and kernel density estimation. [8 marks]

10. In a radio-tracking study, 36 adult canvasback ducks, *Aythya valisineria*, received radio collars. The numbers recovered dead at each of 6 subsequent capture times are given below

Number	Capture time					
Released	1	2	3	4	5	6
36	3	2	3	3	2	1

Suppose that the probability of any duck surviving from one time to the next is ϕ.

(i) Write down an expression for the probability that a duck dies at the jth capture time, $j = 1, 2, \ldots, 6$ and an expression for the probability that a duck survives beyond the last capture time. [5 marks]

(ii) If x_j ducks die at the jth time, $j = 1, 2, \ldots, 6$, and x_7 ducks survive beyond the last capture time, write down an expression for the likelihood in terms of ϕ. [4 marks]

(iii) Verify that the beta distribution is a conjugate prior, and write down the expression for the posterior distribution of ϕ. [4 marks]

(iv) For the data above, and a uniform prior distribution, show that the posterior mean is 0.916, and the maximum-likelihood estimate of ϕ is 0.921. [Note that if a random variable X has a beta, Be(m, n) distribution then its expectation is $m/(m + n)$.] [5 marks]

(v) Explain the relationship between the posterior mode and the maximum-likelihood estimate when the prior distribution is uniform. [2 marks]

(vi) Explain how to obtain a random sample from the posterior distribution for ϕ using rejection, as illustrated in the figure below. [5 marks]

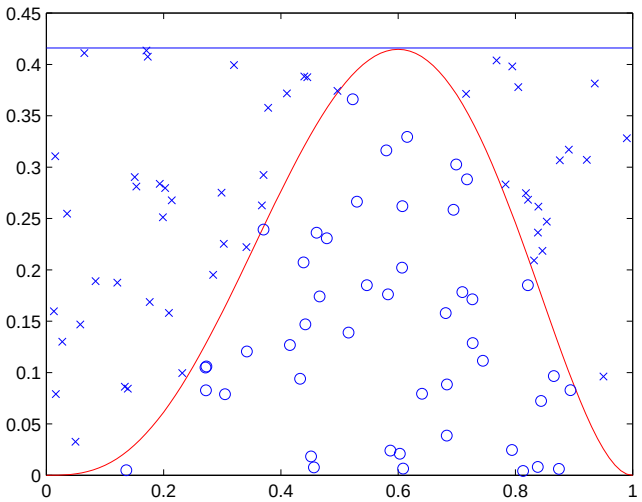

Figure 5: Figure for Question 10.

(vii) As an alternative, explain how to simulate from the posterior distribution for ϕ using a Metropolis Hastings method, to result in a plot like that in Figure 6. [5 marks]

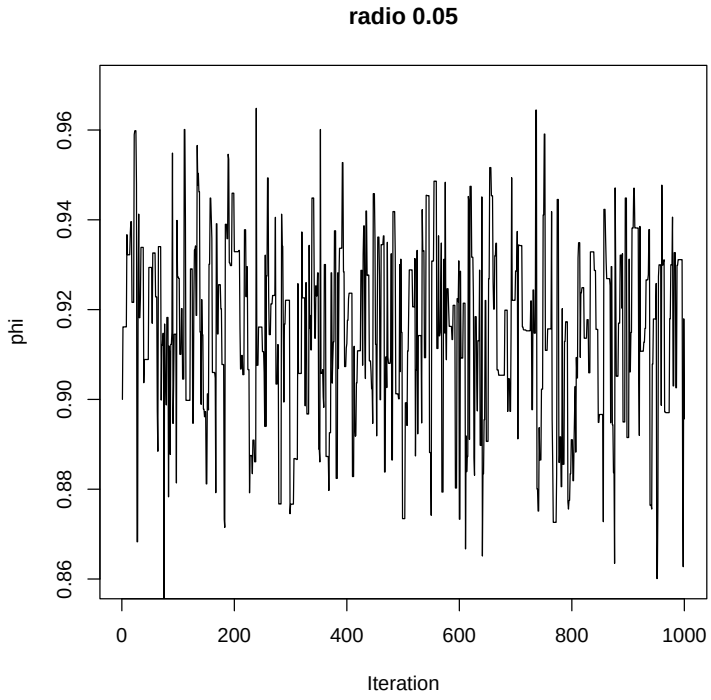

Figure 6: Figure for Question 10.