UNIVERSITY OF KENT

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

LEVEL H EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

Friday, 12 May 2006: 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.*

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. Explain when you would expect to use the Poisson distribution as a model for a univariate random variable. Write short notes, including illustrations, on THREE of the following modifications of a Poisson distribution:

   (i) A Poisson process with one rate until the first event and then a different rate for all subsequent events.

   (ii) A mixture of Poisson distributions.

   (iii) A zero-inflated Poisson distribution.

   (iv) A zero-truncated Poisson distribution. [10 marks]

2. Write short notes on the use of MATLAB programs in statistics. Include in your answer discussion of:

   (i) The difference between script and function files.

   (ii) The use of elementwise matrix operations.

   (iii) The use of standard vector multiplication to form a log-likelihood. [10 marks]

3. The random sample, (0.11, 1.67, 1.01, -1.20, -2.80, -0.68, 2.28, -1.14, -7.34, -4.60) results from the one-parameter Cauchy distribution, with probability density function given by

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

   The maximum-likelihood estimate of θ is $\hat{\theta} = -0.6632$, and Fisher's expected information is $J = 5$.

   Let $L(\theta)$ denote the likelihood. The hypothesis $\theta = 0$ may be tested by means of the score, Wald and likelihood-ratio tests. The results of these tests are shown below:

   (i) $\sqrt{2log\left\{\frac{L(-0.6632)}{L(0)}\right\}} = 1.026$

   (ii) $(0.6632)\sqrt{5} = 1.483$

   (iii) $\frac{\sum_{i=1}^{10}\left\{\frac{dlogf(x_i)}{d\theta}\right\}\Big|_{\theta=0}}{\sqrt{5}} = -0.620.$

   Explain, giving your reasons, which test has been used in each of these cases. [10 marks]

4. On each of T separate visits to N sites it is recorded whether or not a particular bird species is present or absent. An illustration of the data that may result from such a study is given below, for the American blue jay, *Cyanocitta cristata*. Here $T = 11$ and $N = 50$.

	Number of detections											
Species	0	1	2	3	4	5	6	7	8	9	10	11
Jay	17	9	11	6	5	2	0	0	0	0	0	0

For any site, let Y denote the number of visits on which a jay was detected. A particular model for detection gives the following expressions for the probability distribution of Y:

$$Pr(Y = 0) = (1 - \psi) + \psi p^T$$
$$Pr(Y = i) = \psi \binom{T}{i} p^{T-i}(1 - p)^i \quad \text{for} \quad 1 \leq i \leq T.$$

The data recorded are the numbers of sites with $0, 1, 2, \ldots, T$ detections, denoted $\{m_y, y = 0, 1, \ldots, T\}$. Here we have $m_0 = 17, m_1 = 9$, etc. Write down a general expression for the likelihood as a function of ψ and p. Explain how the EM algorithm would operate in this example. [10 marks]

5. An alternative approach to fitting the model of Question 4 is to reparameterise. Show that when the model is reparameterised in terms of $\theta = \psi(1 - p^T)$ and p then there exists an explicit maximum-likelihood estimator for θ.

If you have estimated θ and p, and have expressions for their standard errors, explain how you would use the delta-method to obtain the standard error of ψ. [10 marks]

6. The information in the table below can be used to compare the 4 models, which are all applied to the same data set; standard notation is used. Explain how the AIC is calculated. Evaluate the missing numbers of model parameters in the "d" column. Compare and contrast the relative fits of the models to the data, examined by means of the AIC values, with the measures of absolute fit that result from the Pearson X^2 values. [10 marks]

Model	d	$X^2(df)$	AIC	-2ℓ
1	*	224.12 (10)	5256	5252
2	*	20.69 (8)	5104	5096
3	*	43.91 (9)	5124	5118
4	*	6.03 (7)	5092	5082

SECTION B

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

7. (a) The data in the table below give the numbers of fertility cycles to conception for 100 couples, together with the expected numbers from two models. One of these is the geometric model, and the other is the beta-geometric model. By only considering the ratios of successive expected numbers, deduce which set of expected values comes from which model, giving your reasons.

Cycle	Observed	Expected	
		Model 1	Model 2
1	29	27.5	22.4
2	16	18.3	17.4
3	17	12.6	13.5
4	4	9.0	10.5
5	3	6.6	8.1
6	9	4.9	6.3
7	4	3.8	4.9
8	5	2.9	3.8
9	1	2.3	2.9
10	1	1.8	2.3
11	1	1.5	1.8
12	3	1.3	1.4
> 12	7	7.6	4.8
Total	100		

[11 marks]

(b) On each of k occasions, a closed population of wild animals, of fixed but unknown size N, is sampled at random and any animals in the sample that have not been marked previously are marked and returned to the population. The objective is to estimate N.

The general form for the likelihood is

$$L(N, \boldsymbol{\eta}) \propto \binom{N}{D} \prod_{j=0}^{k} p_j^{f_j},$$

where f_j denotes the number of distinct animals that have been captured j times and D is the number of distinct animals caught, given by $D = \sum_{j=1}^{k} f_j$, so that $f_0 = N - D$. The probability p_j denotes the probability that an animal is caught j times out of the k occasions, and the vector $\boldsymbol{\eta}$ denotes the set of model parameters, excluding N.

Question continued on opposite page

(i) If all animals share the same recapture probability p at each occasion, events at different occasions are independent and animals are caught independently of one another, write down an expression for p_j. [8 marks]

(ii) An alternative model allows the recapture probability to vary between animals. In this case the probability of an animal being caught j times has the expression given below.

$$p_j \propto \frac{\prod_{r=0}^{j-1}(\mu + r\theta)\prod_{r=0}^{k-j-1}(1-\mu+r\theta)}{\prod_{r=0}^{k-1}(1+r\theta)}.$$

Explain in outline only how you think this expression is derived, and provide an interpretation of the two model parameters. [11 marks]

8. (a) The random variable X has a Weibull distribution, with probability density function

$$f(x) = \frac{\beta}{\gamma^\beta} x^{\beta-1} exp\{-(x/\gamma)^\beta\}$$

for $0 \leq x < \infty$ and $\beta > 0, \gamma > 0$.

Construct the cumulative distribution function for X and use this to derive an algorithm for simulating from X. [10 marks]

(b) Consider the case of a random variable X which has a $N(\mu, \sigma^2)$ distribution, where σ^2 is known, and the parameter μ has a $N(\nu, \tau^2)$ distribution. The posterior distribution of μ, corresponding to a sample x of size 1, has mean $\frac{\frac{x}{\sigma^2} + \frac{*}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}$ and variance $\left(\frac{1}{\sigma^2} + *\right)^{-1}$.

(i) Complete the specification of these two terms by evaluating the $*$ components. Discuss whether the prior distribution is conjugate in this example. [12 marks]

(ii) An important issue in Bayesian analysis is the sensitivity of posterior distributions to assumptions made regarding prior distributions. Let the posterior distribution for μ have mean ν^*. One way to evaluate the sensitivity of the posterior distribution to the prior is to form the derivative, $\frac{\partial \nu^*}{\partial \nu}$.

For the example of this question, show that the above derivative is the ratio of posterior variance to prior variance, and discuss whether you think this is a sensible result. [8 marks]

9. Sharon Kennedy conducted a research project in Trinidad in July 1999 on the occurrence and distribution of larvae of the hoverfly *Diptera* on the flowering bracts of *Heliconia* plants. In a small part of her study she observed that on 20 randomly selected bracts the numbers of larvae were

$$9 \quad 14 \quad 3 \quad 3 \quad 8 \quad 7 \quad 7 \quad 6 \quad 7 \quad 0 \quad 6 \quad 0 \quad 5 \quad 1 \quad 3 \quad 12 \quad 2 \quad 4 \quad 0 \quad 11.$$

Let X denote the random variable which describes the number of larvae on a randomly selected bract. Assume that X has the probability function,

$$\Pr(X = x) = \left(\frac{1}{1+\mu}\right)\left(\frac{\mu}{1+\mu}\right)^x, \quad \text{for} \quad x = 0, 1, 2, \ldots,$$

where $\mu > 0$ is the mean value of X.

(a) Show that the log-likelihood function for μ is given by

$$\ell(\mu) = 108 \log(\mu) - 128 \log(1+\mu).$$

[7 marks]

(b) Find the maximum likelihood estimate of μ and also the observed Fisher information. [10 marks]

(c) A plot of the log likelihood function for μ is shown in Figure 1 on the opposite page. Use this graph to obtain an approximate 95% confidence interval for μ. [7 marks]

(d) Compare this interval with an alternative 95% confidence interval for μ which is calculated from using the asymptotic normality of maximum-likelihood estimators. Which would you prefer? [6 marks]

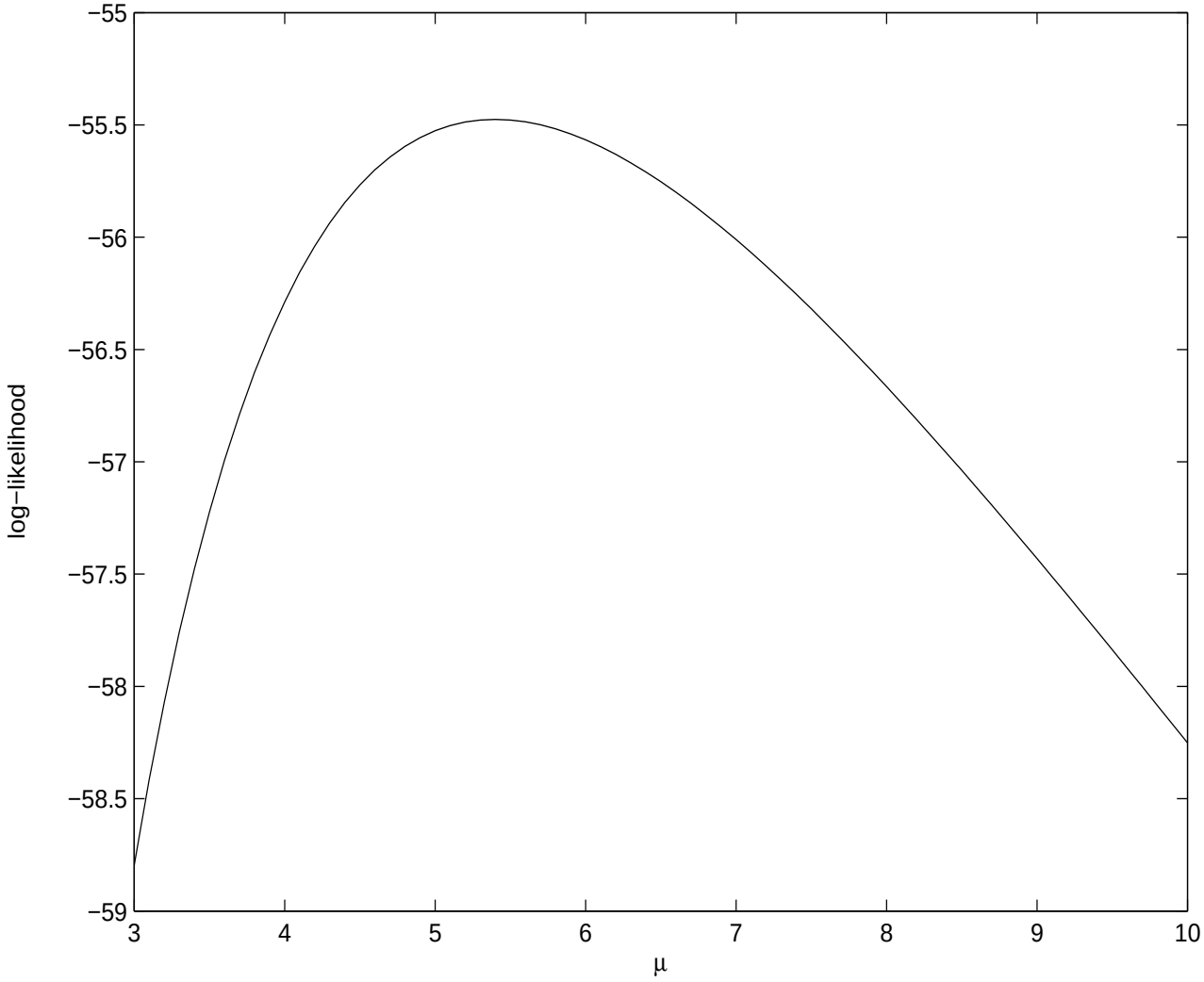

Figure 1: Log-likelihood for Question 9.

10. (a) The MATLAB programs below can be used to maximise the log-likelihood for a two-parameter Cauchy distribution. Shown in Figure 2 are the iterations taken by running the program in the script file.

```
gr=1; global data start gr
data=[1.09   -0.23    0.79    2.31   -0.81]
start=[.5 .5]
while norm(gr)>.0001
  gr=grad('cauchy', start)
  z=fminbnd('linesearch', -0.5,0.5)
  start=start+z*gr'
end

function w=linesearch(x)
global data start gr
w=cauchy(start+x*gr');
```

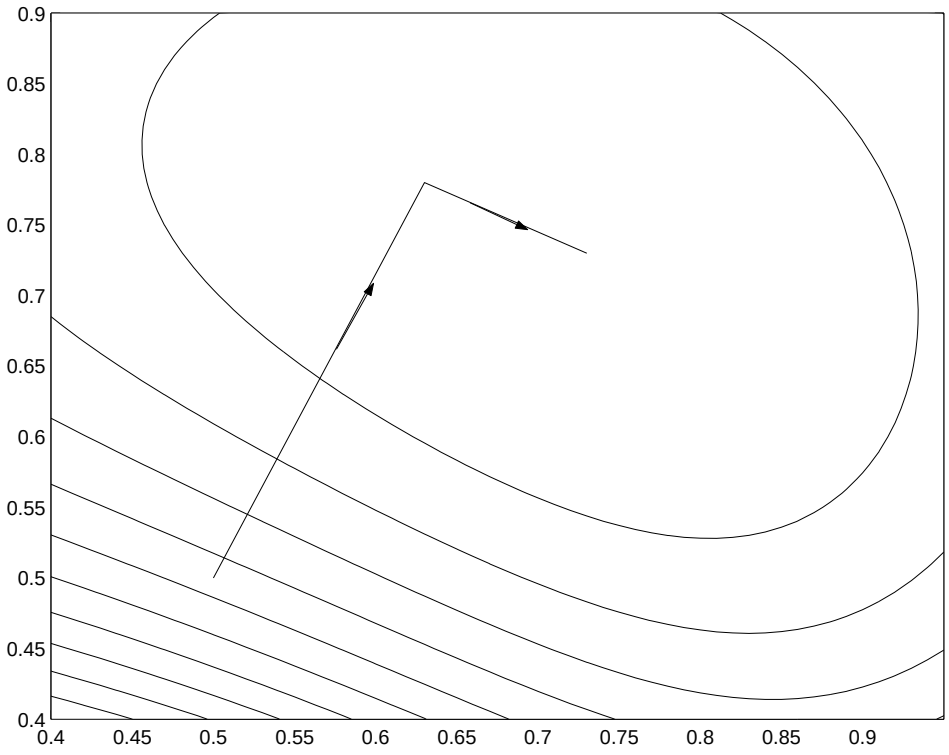

Figure 2: Two-parameter Cauchy log-likelihood.

Explain, giving your reasons, which method is being used. [11 marks]

Question continued on opposite page

(b) Shown in Figure 3 is the progression of the simplex method for seeking the maximum of the same Cauchy log-likelihood surface. The numbers correspond to successive corners of the simplex, as they are added to the iteration. Explain, with reference to the surface contours, how the method is proceeding. [12 marks]

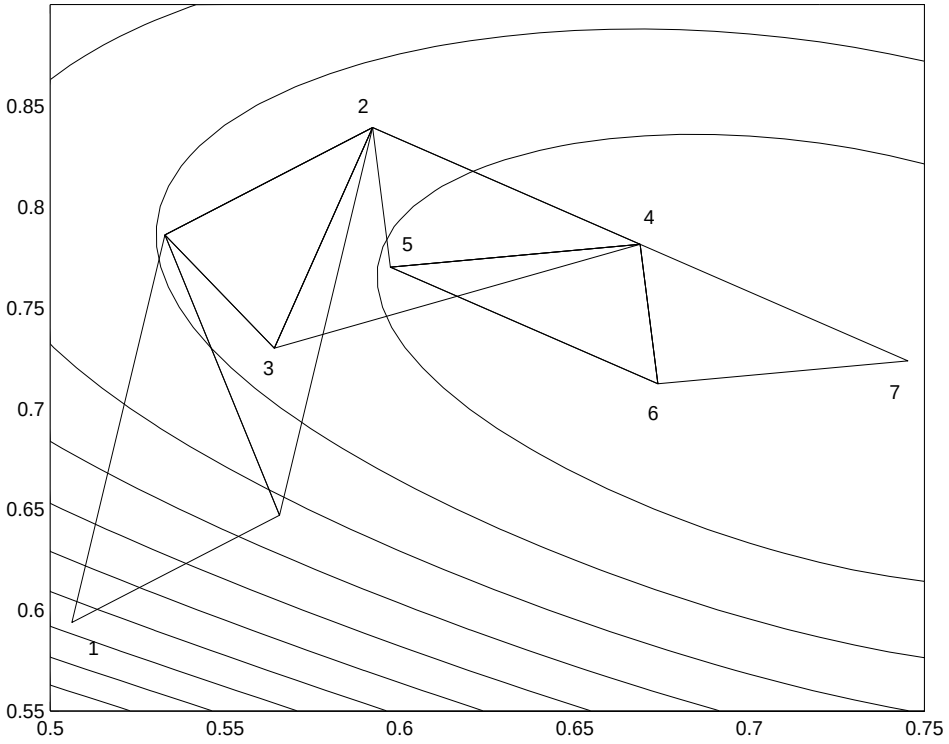

Figure 3: Two-parameter Cauchy log-likelihood.

(c) Discuss the use of surface derivatives in function optimisation. Include in your answer reference to the two examples of this question. [7 marks]