UNIVERSITY OF KENT

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

LEVEL H EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

Saturday, 17 May 2008: 9.30 – 11.30

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. The data below give the numbers of cycles to conception for 486 human couples seeking to conceive.

cycle	1	2	3	4	5	6	7	8	9	10	11	12	>12
number	198	107	55	38	18	22	7	9	5	3	6	6	12

(i) Specify the geometric model for these data. [3 marks]

(ii) Show that the probability, p_k, of conception taking k cycles, has the form shown below when the probability of conception per cycle, p, has a beta distribution,

$$p_k = \frac{B(\alpha+1, \beta+k-1)}{B(\alpha,\beta)},$$

where $B(\alpha,\beta)$ is the beta function.

Note that the beta probability density function is given by

$$f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \quad \text{for} \quad 0 \le p \le 1.$$

[7 marks]

2. Surveys of the blue-ridged two-lined salamander, *Eurycea wilderae*, take place in the Great Smoky Mountains National Park in America. On each of 5 occasions, 39 different sites were surveyed. Thus a record of 01101 at a site would indicate that the salamander was present at the site, with probability Ψ, that it was not detected on the first occasion with probability $1-p_1$, was detected on the second and third occasions, with probabilities p_2 and p_3 respectively, was not detected on the fourth occasion, with probability $1-p_4$, and was detected on the last occasion, with probability p_5. The record would then have probability $\Psi(1-p_1)p_2p_3(1-p_4)p_5$.

(i) Write down the probability of the record 00000. [5 marks]

(ii) The model with constant probability of detection, p, not varying with time, has an Akaike information criterion that is 1.95 less than that for the model above. Explain how the measure is formed, and discuss which model you would prefer to use.

[5 marks]

3. A batch of n eggs is infected by bacteria, *Salmonella enteriditis*, and the number of bacteria per egg is thought to be described by a Poisson distribution with mean μ. It is observed that r eggs are free from infection.

   (i) Write down the likelihood. [4 marks]

   (ii) Obtain the maximum-likelihood estimator for μ and show that its variance is approximately given by

$$Var(\hat{\mu}) \approx \frac{(e^{\hat{\mu}} - 1)}{n}.$$

   [6 marks]

4. Extreme climate events result in data that may be modelled by the generalised extreme value distribution, which has cumulative distribution function given by

$$F(x) = \exp\left[-\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi}\right],$$

$$\text{for } 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0, \text{ and } -\infty < \mu, \xi < \infty, \quad \sigma > 0.$$

   (i) Write down an expression for the probability density function of the generalised extreme value distribution. [2 marks]

   (ii) Hence explain how you would form the likelihood when there is a random sample of n known values and a corresponding sample of m right-censored values. [2 marks]

   (iii) Annual maximum sea levels at Port Pirie in South Australia for the period 1923 – 1987 result in maximum-likelihood estimates given by

$$\hat{\mu} = 3.87, \quad \hat{\sigma} = 0.198, \quad \hat{\xi} = -0.050.$$

   By the usual procedures, the estimated variance covariance matrix is given by

$$V = \begin{bmatrix} 0.000780 & 0.000197 & -0.001070 \\ 0.000197 & 0.000410 & -0.000778 \\ -0.001070 & -0.000778 & 0.009650 \end{bmatrix},$$

   where the order of rows and columns corresponds to μ, σ, ξ.

   Explain how you would proceed to obtain such a matrix V. [4 marks]

   (iv) Stating the asymptotic distribution for maximum-likelihood estimators, obtain approximate 95% confidence intervals for each of the three parameters. [2 marks]

5. In question 3, if p denotes the probability of an egg escaping infection, write down an expression for the variance of the maximum-likelihood estimator, $\hat{p}$. Use the delta method to approximate the variance of $\hat{\mu}$, and compare the result with the answer to part (ii) of the last question. [10 marks]

6. In order to simulate random variables from the gamma distribution with probability density function

$$f(x) = \frac{x^{n-1}e^{-x}}{\Gamma(n)}, \quad \text{for} \quad x \geq 0 \quad \text{and} \quad n > 1,$$

a rejection method is to be used, with the enveloping function given by an appropriate multiple of the exponential density,

$$g(x) = ke^{-x/n}/n, \quad \text{for} \quad x \geq 0, \quad \text{and suitable} \quad k > 1.$$

Describe the rejection algorithm. (You do not need to derive an expression for k). [10 marks]

SECTION B

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

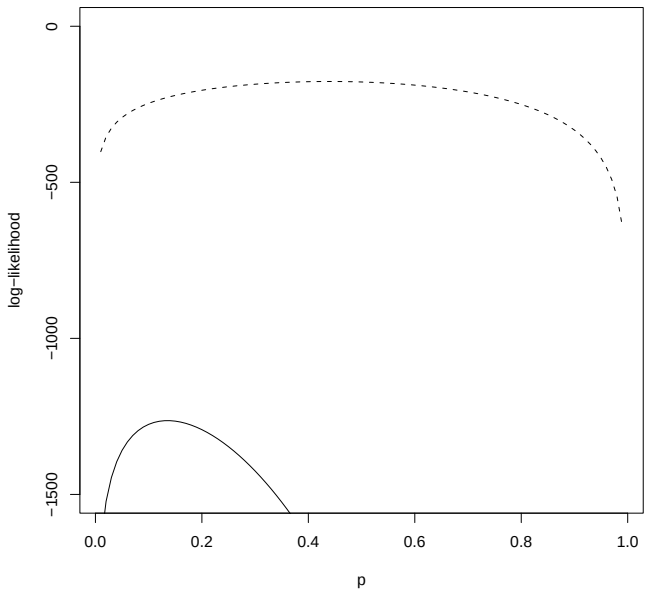

Figure 1: Figure for Question 7.

7. (a) In a Schnabel census of marked wild animals, data $\{f_i, 1 \le i \le t\}$ are collected, where f_i animals are caught i times and t is the number of sampling occasions. Two real illustrations of $\{f_i\}$ are shown below. For the voles, $t = 5$, while for the golftees, $t = 8$.

i	1	2	3	4	5	6	7	8
voles	18	15	8	6	5	–	–	–
golftees	46	28	21	13	23	14	6	11

If p denotes the capture probability, and $\overline{p} = 1 - p$, then the likelihood has the form,

$$L(p|\boldsymbol{f}) \propto \prod_{i=1}^{t} \left\{ \frac{\binom{t}{i} p^i \overline{p}^{t-i}}{1 - \overline{p}^t} \right\}^{f_i}. \tag{1}$$

Graphs of $\ell(p|\boldsymbol{f}) = \text{logL(p}|\boldsymbol{f})$ are given in Figure 1, for the two examples.

(i) Explain how the expression of equation (1) is derived. [6 marks]

(ii) Discuss the different forms of the two log-likelihoods, indicating which you think corresponds to which data set. [4 marks]

(iii) Explain in outline only how you would use these graphs to obtain likelihood-based confidence intervals for p. [5 marks]

(b) The data below give the distribution of stillbirths in 402 litters of New Zealand white rabbits.

No. of stillbirths	0	1	2	3	4	5	6	7	8	9	10	11
No. of litters	314	48	20	7	5	2	2	1	2	0	0	1

For data of this kind one might want to fit a zero-inflated Poisson distribution, with probabilities given by

$$\begin{aligned} \text{pr}(Y = 0) &= \omega + (1 - \omega)p_0 \\ \text{pr}(Y = i) &= (1 - \omega)p_i, \quad \text{for} \quad i = 1, 2, \ldots \end{aligned}$$

where the random variable Y denotes the number of stillbirths, $\omega(> 0)$ is a zero-inflation probability and $\{p_i\}$ is the Poisson distribution, with mean λ. A score test of the null hypothesis that $\omega = 0$ has test statistic of the form

$$z = \boldsymbol{U}' \boldsymbol{J}^{-1} \boldsymbol{U},$$

where $\boldsymbol{U}$ is the scores vector containing the partial derivatives of the log-likelihood with respect to ω and λ, evaluated when $\omega = 0$ and $\lambda = \hat{\lambda}_0$, where $\hat{\lambda}_0$ is the maximum likelihood estimate of λ when $\omega = 0$.

(i) Obtain $\hat{\lambda}_0$. [5 marks]

(ii) Define the matrix $\boldsymbol{J}$. [5 marks]

(iii) For the rabbit data, the score test statistic has the value $z = 182.67$. Discuss this result in the light of the data, and compare the conclusions of the score and likelihood ratio tests. Note that maximised log-likelihood values are given as $\ell(\hat{\omega}, \hat{\lambda}) = -357.1892$ and $\ell(0, \hat{\lambda}_0) = -440.8435$. [5 marks]

8. Three MATLAB programs are given below. One of these is a function that specifies the Bohachevsky surface, which possesses a number of optima, and may therefore be used to illustrate the performance of numerical optimsation procedures. The other two MATLAB programs produce the scatter plots of Figures 2 and 3, that are shown opposite.

(a)

```
for i=1:10000
 x=(2*(rand(1,2)-0.5));
 w=fminsearch('boha_surface', x);
  if(abs(w)<0.001)
  plot(x(1),x(2),'.')
  else
end
end
```

(b)

```
function z=boha(w)
x=w(1);
y=w(2);
z=x.*x+2*y.*y-0.3*cos(3*pi*x)-0.4*cos(4*pi*y)+0.7;
```

(c)

```
for i=1:20
 x=(2*(rand(1,2)-0.5));
 w=annealing_boha('boha_surface',x);
 plot(w(1),w(2),'.')
end
```

(i) Explain which program you think produces which figure. What is the name of the file containing the second program? [10 marks]

(ii) The optimisation methods used in two of the programs are the simplex method and simulated annealing. Explain which program uses which method, and explain in detail how each of these programs and methods work. [13 marks]

(iii) What can you conclude from the scatter plots of the two figures? [7 marks]

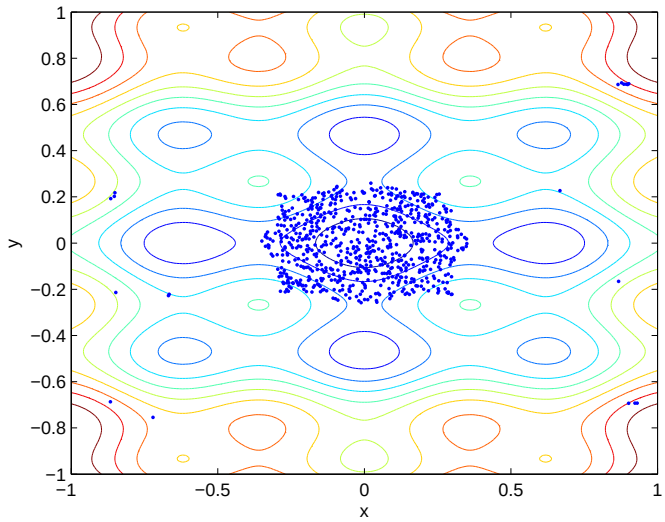

Figure 2: Figure for Question 8.

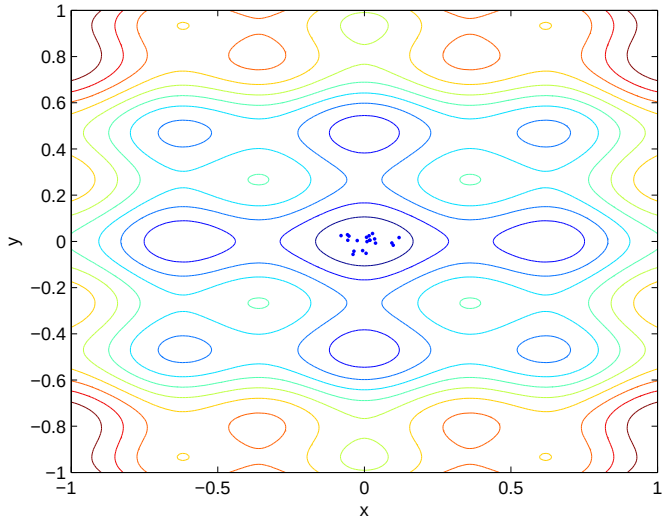

Figure 3: Figure for Question 8.

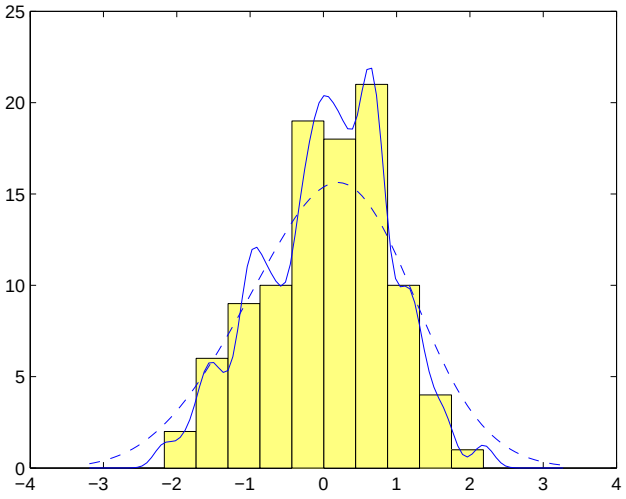

Figure 4: Figure for Question 9.

9. (a) Shown in Figure 4 is a histogram of 100 $N(0,1)$ random variates, and two kernel density estimates, resulting from using a normal kernel. Explain how the kernel density estimates are formed, and how they can differ. Discuss which you would prefer in this illustration.

[16 marks]

(b) The space shuttle Challenger exploded after its launch on the 28^{th} January, 1986. It was thought that rubber insulating rings, called 'O-rings', had lost their flexibility due to the cold of the night before the launch. Data from previous flights allowed the collection of information on the number of O-rings out of 6 that were so affected, at each of a number of different temperatures.

Describe how you would analyse such data, and suggest appropriate models that you might fit to the data. [14 marks]

10. In a genetic linkage study, 197 animals are divided into four categories with respective probabilities given by

$$\left(\frac{1}{2}+\frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

(i) Explain how the EM algorithm may be used to obtain the maximum likelihood estimate of θ. [13 marks]

(ii) With a uniform prior density for θ, write down the posterior density for θ, ignoring the constant of proportionality. [4 marks]

(iii) Describe how you would sample from the posterior density using the Metropolis-Hastings algorithm. [13 marks]