MA771/01

UNIVERSITY OF KENT AT CANTERBURY

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

PART II EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

Tuesday, 1st May 2001: 2.00 p.m. to 4.00 p.m.

This paper contains FIVE *questions. Candidates should not attempt more than* THREE *questions.*

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

Each question will be marked out of 40.

Turn over

1. (a) When introduced into an area containing both fine and coarse sand, ant-lions are thought to prefer to dig burrows in fine sand, but also to avoid other ant-lions. The data below describe the results of 62 experiments in which either 3 or 4 ant-lions are introduced into an area with fine and coarse sand. Explain how you would describe the data by means of a binomial model with probability p of burrowing in fine sand.

Number of ant-lions introduced	Total number of experiments	Number of ant-lions burrowing in fine sand				
		0	1	2	3	4
3	32	0	7	24	1	
4	30	0	3	17	10	0

Obtain the maximum-likelihood estimate $\widehat{p}$, and use it to form expected values corresponding to the observed values in the above table.

Without formally computing a goodness-of-fit statistic, discuss the goodness-of-fit of the binomial model to the data by visually comparing observed and expected cell numbers in the above table. [26 marks]

(b) A species of wasp lays its eggs on larvae. Suppose that any larva has X encounters with wasps where X has the Poisson distribution,

$$\Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad \text{for} \quad k = 0, 1, 2, \ldots$$

Suppose that an egg is always laid at the first encounter, but that at all subsequent encounters, single eggs are laid independently with probability $\delta < 1$ at each encounter. If p_r denotes the probability that a larva receives r eggs, $(r = 1, 2, \ldots)$ then

$$p_r = \delta^{r-1}(\delta - 1)^{-r} e^{-\lambda\delta} \sum_{i=r}^{\infty} \frac{\lambda^i(\delta - 1)^i}{i!}.$$

Verify this for $r = 1$ and $r = 2$. [14 marks]

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

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

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

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

[14 marks]

(ii) Fisher's expected information is $J = \frac{n}{2}$. Write down the corresponding iteration for the method of scoring. [7 marks]

(iii) A particular sample has the values: 0.11, 1.67, 1.01, –1.20, –2.80, –0.68, 2.28, –1.14, –7.34, –4,60, resulting in $\widehat{\theta} = -0.6632$.

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:

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

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

(c) $\frac{\sum_{i=1}^{10} \left\{ \frac{d \log f(x_i)}{d\theta} \right\}\Big|_{\theta=0}}{\sqrt{5}} = 0.620$

Explain, giving your reasons, which test is used in each of these cases. Comment on the disparity between the three test results. [19 marks]

3. In the ABO blood group system there are four blood groups, A, B, AB and O, occuring with respective relative frequencies, $(p^2 + 2pr)$, $(q^2 + 2qr)$, $2pq$ and r^2, where p, q and r are probabilities and $p + q + r = 1$.

(i) Write down the likelihood, L, when the different groups are observed with frequencies n_A, n_B, n_{AB} and n_0 respectively, with $n = n_A + n_B + n_{AB} + n_0$. [8 marks]

(ii) Writing $r = 1 - p - q$, and treating L as a function of p and q, show that

$$\frac{\partial \log L}{\partial p} = \frac{2r}{p(p+2r)} n_A - \frac{2}{(q+2r)} n_B + \frac{1}{p} n_{AB} - \frac{2}{r} n_0.$$

[9 marks]

(iii) Given that

$$-\frac{\partial^2 \log L}{\partial p \partial q} = \frac{2}{(p+2r)^2} n_A + \frac{2}{(q+2r)^2} n_B + \frac{2}{r^2} n_0,$$

show that

$$-\mathbb{E}\left[\frac{\partial^2 \log L}{\partial p \partial q}\right] = \frac{2n(4r+3pq)}{(p+2r)(q+2r)}.$$

[10 marks]

(iv) An iterative method for maximising the likelihood is called the *gene-counting* method. Here

$$\widehat{p}^{(i+1)} = \left(\frac{n_A + n_{AB}}{2n}\right) + \frac{\widehat{p}^{(i)}}{(\widehat{p}^{(i)} + 2\widehat{r}^{(i)})} \frac{n_A}{2n}$$

$$\widehat{q}^{(i+1)} = \left(\frac{n_B + n_{AB}}{2n}\right) + \frac{\widehat{q}^{(i)}}{(\widehat{q}^{(i)} + 2\widehat{r}^{(i)})} \frac{n_B}{2n},$$

where $\widehat{q}^{(m)}$ and $\widehat{p}^{(m)}$ are, respectively, the mth iterates of the maximum-likelihood estimates, $\widehat{q}$ and $\widehat{p}$. By separately writing each of the frequencies n_A and n_B as the sum of two quantities, one of which is missing, explain how gene-counting arises from applying the EM algorithm. [13 marks]

4. (a) Describe how to simulate random variables from an exponential distribution with probability density function

$$f(x) = \lambda e^{-\lambda x} \quad \text{for} \quad x \geq 0,$$

using the inversion method. [8 marks]

(b) A $\Gamma(n, 1)$ random variable has the probability density function

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

Describe how to simulate such random variables, using the rejection method with envelope

$$g(x) = \frac{ke^{-x/n}}{n}, \quad \text{for} \quad x \geq 0$$

and suitable $k > 1$. [13 marks]

(c) Data, $\mathbf{x} = (x_1, x_2, x_3)$ arise with the multinomial probability,

$$f(\mathbf{x}|\theta, \eta) \propto \theta^{x_1}\eta^{x_2}(1 - \theta - \eta)^{x_3},$$

for $0 < \theta,\ \eta < 1$ and $\theta + \eta < 1$.

The joint prior distribution for the pair of parameters (θ, η) is said to be *Dirichlet*, with probability density function

$$\pi(\theta, \eta) \propto \theta^{\alpha_1}\eta^{\alpha_2}(1 - \theta - \eta)^{\alpha_3},$$

for given values of α_1, α_2 and α_3.

Show that the posterior distribution $\pi(\theta, \eta|\mathbf{x})$ is also Dirichlet. Find the conditional posterior distributions for θ and η, each given the other, and outline how these might be used in Gibbs sampling. [19 marks]

5. (a) A random sample $x_1, \ldots, x_n$ is taken from a probability density function, $f(x)$. Define the *naive estimate* of $f(x)$. Explain how the naive estimate differs from a standard histogram. Explain how the naive estimate of $f(x)$ is a kernel density estimate of $f(x)$. [21 marks]

(b) 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 role played by *delta*, and suggest two possible forms 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);
```

[13 marks]

(c) In Monte Carlo inference, parameters are estimated by maximising an approximate likelihood, obtained from using simulated data. Explain the rôle played by kernel density estimation in Monte Carlo inference. [6 marks]