UNIVERSITY OF KENT

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

PART II EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

Thursday, 1st May 2003: 9.30 a.m. to 11.30 a.m.

This paper contains FIVE *questions. Candidates must not attempt more than* THREE *questions. All questions will be marked out of 40.*

Copies of the New Cambridge Statistical Tables *are provided. Graph paper is available. Approved calculators may be used.*

1. The data below provide the number of cycles to conception for couples successful in conceiving. The data are classified according to whether the female of the couple smoked, and are censored at 12 cycles.

Cycle	Women smokers	Women non-smokers
1	29	198
2	16	107
3	17	55
4	4	38
5	3	18
6	9	22
7	4	7
8	5	9
9	1	5
10	1	3
11	1	6
12	3	6
>12	7	12
Total	100	486

If the random variable X denotes the number of cycles to conception for any couple, the *geometric* model for the data is given by $\Pr(X = k) = p(1-p)^{k-1}$ for $k = 1, 2, \ldots$

(a) Interpret the parameter p. Write down the likelihood for a single data set censored at r cycles, with $\{n_i,\ i = 1, 2, \ldots, r\}$ individuals waiting i cycles, and n_{r+1} individuals having a waiting time $> r$ cycles. [10 marks]

(b) Show that the maximum-likelihood estimate of p is given by

$$\hat{p} = \sum_{i=1}^{r} n_i \Big/ \left(\sum_{i=1}^{r} i n_i + r n_{r+1} \right).$$

Provide an intuitive explanation for this expression, and compute $\hat{p}$ for the couples where the women were smokers. [13 marks]

Question continued on opposite page

(c) An alternative model has the probability function

$$\Pr(X = k) = \frac{\mu \prod_{i=1}^{k-1} \{1 - \mu + (i-1)\theta\}}{\prod_{i=1}^{k} \{1 + (i-1)\theta\}} \qquad \text{for } k = 1, 2, \ldots,$$

where we interpret $\prod_{i=1}^{0} = 1$.

Identify this model, comment briefly on its derivation, and interpret the model parameters. Explain whether you think this model would provide a better fit to the data than the geometric model. [13 marks]

(d) Explain in outline only how you would use both models to produce a statistical test of the effect of a woman smoking on the rate of conception. You are not expected to carry out the tests. [4 marks]

2. (a) A model for the spread of a measles epidemic initiated by one infective in a household of size 3 predicts that the total number of people infected in the household has 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, the probability of adequate contact, is an unknown parameter. The following data give the frequency distribution of 334 such measles epidemics.

Total number infected	1	2	3
Frequency	34	25	275

Obtain a quadratic equation that is satisfied by the maximum-likelihood estimate $\hat{\theta}$, and deduce that $\hat{\theta} = 0{\cdot}728$.

Carry out a Pearson goodness-of-fit test to examine if the model provides a good fit to these data. [20 marks]

(b) A more refined model splits those epidemics in which 3 people are infected into two types: 3(a), in which the initial infective infects both the other individuals in the household and 3(b), in which the initial infective only infects one of the other individuals in the household, who then infects the third individual. The corresponding probability distribution is

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

Suppose that n independent epidemics are observed and their final outcomes, $y_1, y_2, \ldots, y_n$ say, recorded, yielding the following frequency distribution.

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

Derive the posterior distribution $f(\theta \mid y)$ when the prior distribution of θ is the beta distribution, with probability density function

$$f(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \qquad (0 < \theta < 1),$$

where α and β are given, and

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

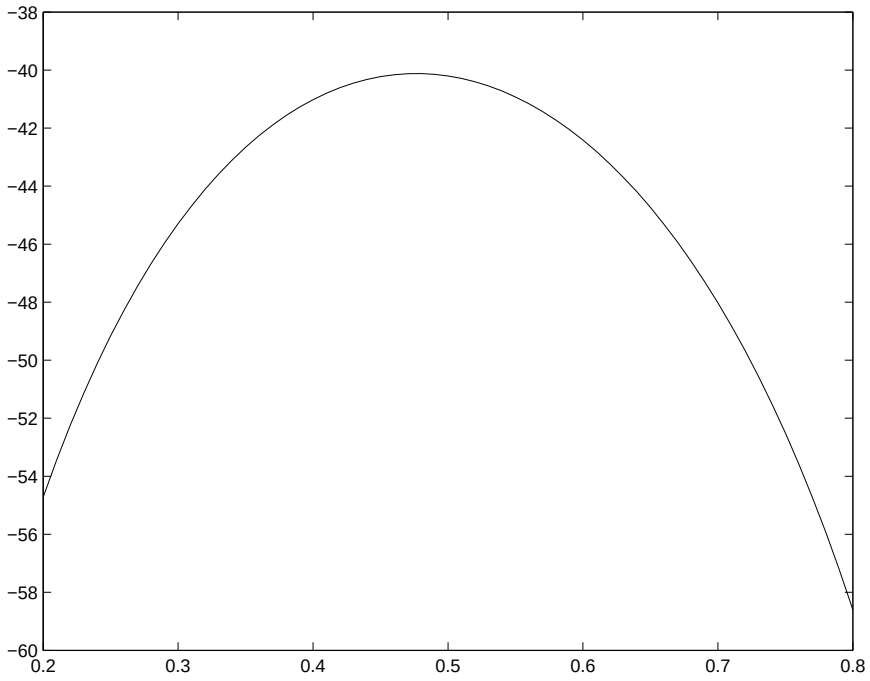
Is the beta distribution a conjugate prior for the case when the final outcomes 3(a) and 3(b) are not separated, as in part (a)? [12 marks]

(c) If the values of c and d are not known, show how the breakdown into types 3(a) and 3(b) in part (b) may be used to provide an EM algorithm for a classical maximisation of the likelihood. You can use the example of part (a) as an illustration, but you are not expected to carry out extensive iterations. [8 marks]

3. (a) For the model of question 2(a), suppose that the total number infected, i, occurs with frequency X_i, $i = 1, 2, 3$. Obtain the expected Fisher information. Write down the asymptotic distribution of the maximum-likelihood estimator $\hat{\theta}$, and use it to obtain an approximate $99\frac{1}{2}\%$ confidence interval for θ when $X_1 = 6$, $X_2 = 11$, $X_3 = 13$, and $\hat{\theta} = 0{\cdot}4765$.

[16 marks]

(b) The graph below presents the plot of the log-likelihood $\ell(\theta)$, for the data of (a) above, against θ. Use this graph to obtain an alternative approximate $99\frac{1}{2}\%$ confidence interval for θ. Comment on the comparison of the two confidence intervals and explain, giving your reasons, which interval you would prefer for general use.



[14 marks]

(c) In standard notation, an iterative procedure for evaluating $\hat{\theta}$ proceeds according to the iteration

$$\theta^{(r+1)} = \theta^{(r)} + d\ell/d\theta \frac{\{\theta^{(r)}(2\theta^{(r)} - 3)\}}{2X_{\cdot}(\theta^{(r)2} - 4\theta^{(r)} - 3)}$$

where $X_{\cdot} = X_1 + X_2 + X_3$. Explain whether this is the Newton-Raphson method, or the method-of-scoring. For the data of part (a), one of these methods produces the sequence

A: $0{\cdot}01 \to 0{\cdot}608 \to 0{\cdot}464 \to 0{\cdot}477 \to 0{\cdot}4765$.

The other produces the sequence

B: $0{\cdot}01 \to 0{\cdot}02 \to 0{\cdot}04 \to 0{\cdot}08 \to 0{\cdot}146 \to 0{\cdot}259 \to 0{\cdot}398 \to 0{\cdot}471 \to 0{\cdot}4765$.

Explain which sequence you think comes from which method (no calculations are required).

[10 marks]

4. (a) The Weibull random variable X has the probability density function (pdf)

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

Write down an expression for the corresponding cumulative distribution function (cdf). Use the cdf to show that we can simulate X by setting

$$X = \frac{1}{\rho}\{-\log_e(U)\}^{1/\rho},$$

where $U \sim U(0,1)$. [11 marks]

(b) Explain in outline the fundamental differences between the computational Bayesian procedures of Gibbs sampling and Metropolis-Hastings. Metropolis-Hastings is to be used to simulate random variables from the Cauchy pdf,

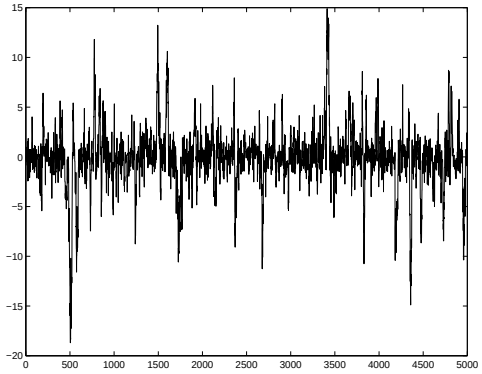
$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

MATLAB code for doing this is given below. Explain the three steps marked with a star (*).

```
strg='1./(1+x.^2)';
cauchy=inline(strg,'x');
strg='1/sig*exp(-0.5*((x-mu)/sig) .^2)';
norm=inline(strg,'x','mu','sig');
n=10000;
sig=2;
x=zeros(1,n);
x(1)=randn(1);% generate the starting point
for i = 2:n
  y=x(i-1)+sig*randn(1);  *
  % generate a uniform for comparison
  u=rand(1);
  alpha=min([1,cauchy(y)*norm(x(i-1),y,sig)/...
        (cauchy(x(i-1))*norm(y,x(i-1),sig))]);  *
  if u <= alpha   *
    x(i)=y;
  else
    x(i)=x(i-1);
  end
end
```

Explain why this method is also a Metropolis method. A trace plot from running the code with $n = 5000$ is shown on the opposite page.

Question continued on opposite page



Explain what is meant by a 'burn-in' period, and suggest how you would select one for this example. [29 marks]

5. (a) The MATLAB function `logit` is given below.

```
function y=logit(t)
global x n r
w=ones(size(x));
alpha=t(1); beta=t(2);
y=w ./(1+exp(alpha+beta*log(x))); z=w-y;
loglik=r*(log(y))'+(n-r)*(log(z))';
y=-loglik;
```

Explain in detail what the data vectors `x`, `n` and `r` represent, and what this function does. Explain how you would use this function in practice in order to obtain maximum-likelihood estimates of the logit model parameters. [12 marks]

(b) The data below describe the mortality of adult flour-beetles (*Tribolium confusum*) after 5 hours' exposure to gaseous carbon disulphide (CS_2).

Dose(CS_2 mg/l)	49·06	52·99	56·91	60·84	64·76	68·69	72·61	76·54
Number of beetles in experiment	59	60	62	56	63	59	62	60
Number of beetles killed	6	13	18	28	52	53	61	59

Define logits, and plot logits vs dose levels. Use this plot to discuss whether you think the logit model is appropriate for the data. [12 marks]

(c) Define the ED_{50}, μ, for an experiment such as that of part (b) above. Provide a rough estimate of μ from the data. Explain how you would obtain an estimate of the variance of $\hat{\mu}$, the maximum likelihood estimator of μ, with and without using the δ-method, for the logit model. [16 marks]