UNIVERSITY OF KENT

FACULTY OF SCIENCE, TECHNOLOGY AND MEDICAL STUDIES

LEVEL H EXAMINATION

APPLIED STOCHASTIC MODELLING AND DATA ANALYSIS

Thursday, 5 May 2005: 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. Compare and contrast the Newton Raphson method and the method-of-scoring for obtaining maximum-likelihood estimates of model parameters when explicit estimates are not available. [10 marks]

2. Describe two methods for obtaining confidence intervals for a scalar parameter θ, when a model is fitted to data by the method of maximum-likelihood. Explain which of the two methods will always produce a symmetric confidence interval, and which will not. Discuss the relative merits of these two methods. [10 marks]

3. An insecticide is being tested for its performance in killing beetles. It is sprayed onto groups of beetles in differing doses, with the ith dose, d_i, applied to a group of size n_i, and with r_i beetles observed to die. In this example the sex of the beetles is not known. Write down a probability model for these data, in terms of the logit function. Explain how a likelihood can be formed. It is often useful to summarise the data by fitting the model by maximum-likelihood and estimating the ED_{50}. Explain what this is, how it is estimated, and whether you think it would provide a useful summary of the beetle experiment. [10 marks]

4. Explain what is meant by the EM algorithm. In your answer provide an illustrative example, giving in detail the different steps in the EM procedure. [10 marks]

5. Explain why the exponential distribution is often used as a simple model for survival times. The Weibull distribution has probability density function given by

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

Explain why its survivor function has the form

$$S(t) = exp\{-(\rho t)^{\kappa}\}.$$

Explain how you would use this expression to simulate values from the Weibull distribution. How does the Weibull model reduce to the exponential distribution? [10 marks]

6. Explain in detail how and why the Gibbs sampler is used. Provide an example to illustrate your answer. [10 marks]

SECTION B

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

7. An investigation was carried out into the effect of artificial playing pitches on the results of certain professional English soccer teams. In particular, the results for Luton Town on their own pitch over the periods 1980-1985 (grass pitch) and 1986-1990 (plastic pitch) are given. The performance of the team is measured as either a Won, Drawn or Lost game.

	Won	Drawn	Lost
Grass	61	31	32
Plastic	65	34	18

It is desired to fit a single-parameter multinomial model to the plastic pitch results with parameter θ and

$$\mathrm{p(win)} = 1 - \theta, \quad \mathrm{p(draw)} = \theta - \theta^2, \quad \mathrm{p(lose)} = \theta^2.$$

(a) If there are observed counts of x_1, x_2, x_3 for the numbers of wins, draws and losses respectively, out of n games, write down the likelihood, and obtain the maximum-likelihood estimate for θ. [10 marks]

(b) Fit the model to the plastic pitch data by maximum likelihood estimation, and carry out a suitable test for goodness of fit. [10 marks]

(c) Fit the same model to the grass pitch results, and to the data for both types of pitch taken together. Calculate the log-likelihood in all three cases (grass, plastic, combined) and hence perform a maximum likelihood ratio test of the hypothesis that $\theta_g = \theta_p$, where θ_g and θ_p are respectively the values of θ for the grass and plastic pitches. [10 marks]

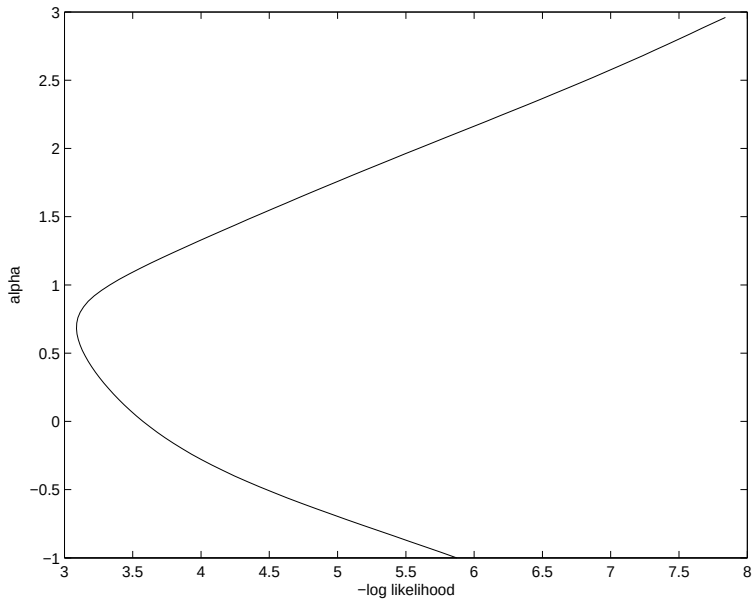

Figure 1: Profile log-likelihood

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

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

(a) 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.

(b) Explain the difference between Figures 1 and 2, and explain the relevance of the path of Figure 3 to the profile of Figure 1. [10 marks]

(c) Describe how you would obtain a confidence interval for α from the profile of Figure 1. [6 marks]

(d) Explain the connection between the log-likelihood section of Figure 2 and kernel density estimation. [8 marks]

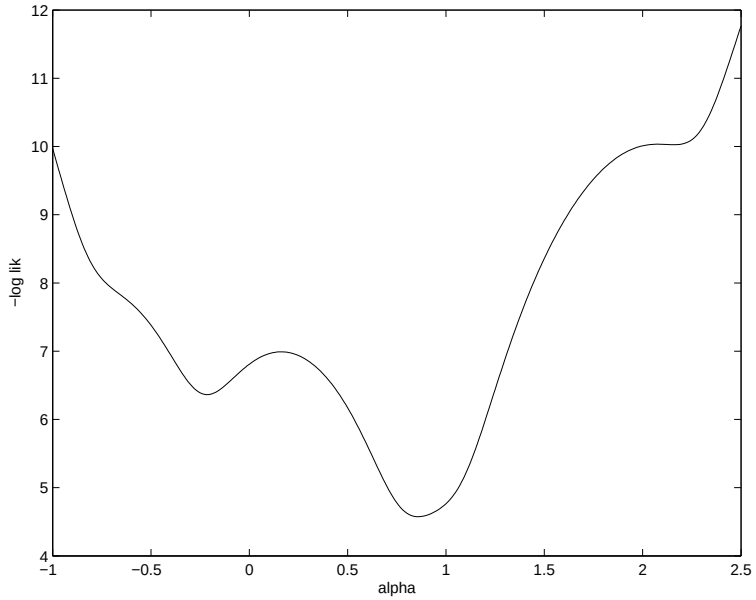

Figure 2: Section of log-likelihood surface, when $\beta = 0.2$

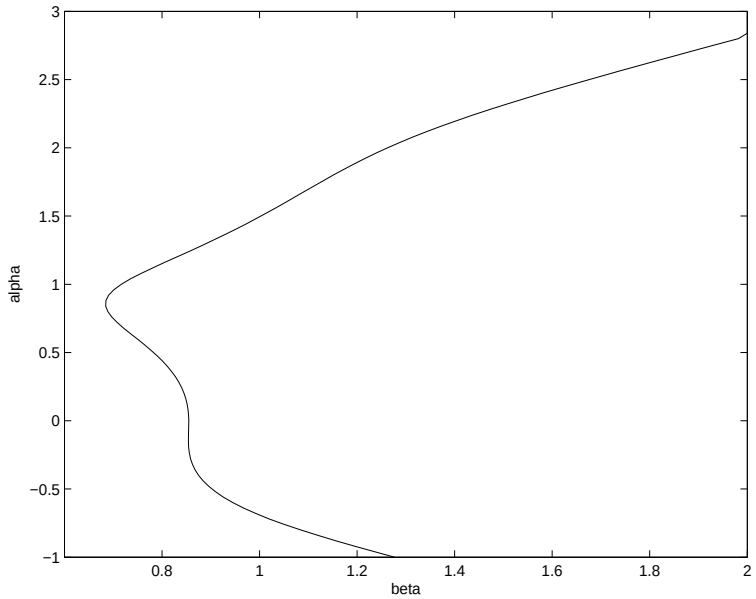

Figure 3: Path traversed in the parameter space corresponding to the profile log-likelihood

9. The data of the table below describe the recaptures of dippers, which are birds found in fast-flowing mountain streams. Each bird was marked before release and it was noted in which year after ringing any bird was first recaptured.

Year of release	Number released	Year of recapture(1981+) 1	2	3	4	5	6
1981	22	11	2	0	0	0	0
1982	60		24	1	0	0	0
1983	78			34	2	0	0
1984	80				45	1	2
1985	88					51	0
1986	98						52

A probability model for these data has two probabilities as parameters, p, the probability of capture in any year and ϕ, the annual dipper survival probability. For the first five rows of the table the data are multinomial, with first two multinomial probabilities respectively, ϕp and $\phi^2 p(1-p)$.

(a) Describe fully the probability distribution for the 80 dippers released in 1984, noting that the study stopped in 1987. [10 marks]

(b) Figure 4 contains several panels. In one is shown the likelihood for this application. In two are shown the marginal posterior distributions for each of the parameters, corresponding to independent $U(0,1)$ prior distributions for the two parameters.

Explain in detail how the marginal distributions are obtained. [10 marks]

(c) In the last two Figure panels are shown histograms resulting from simulations from the two marginal posterior distributions. The simulations have been obtained by means of a Markov chain Monte Carlo procedure, the MATLAB code for which is shown in the program below.

```
m=500; k=100;
for j=1:k
    b=[.9,.4];
    for i=2:m
        a=[rand, rand];
        alpha=min(1,dipper(a)/dipper(b));
        if alpha > rand;
            b=a;
        else
        end
    end
    x(j)=b(1);y(j)=b(2);
end
subplot(3,3,3), hist(y); subplot(3,3,7), hist(x);
```

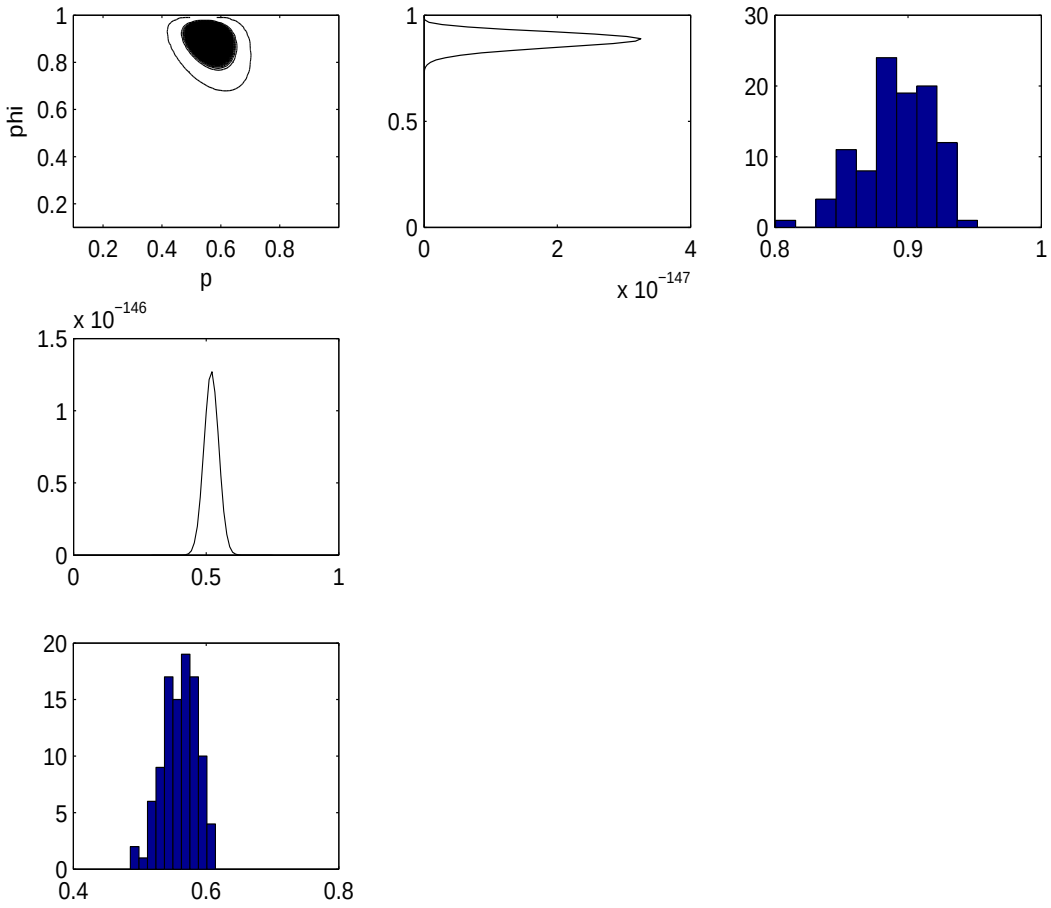

Figure 4: Dipper likelihood, marginals and simulations

Explain why statisticians frequently use Markov chain Monte Carlo approaches, and provide a full description of how the method of the MATLAB program operates. Include in your answer an evaluation of whether you think this Markov chain Monte Carlo procedure might be made more efficient, and of the method used to obtain the samples described by the histograms.

[10 marks]

10. An experiment took place in St. Andrews University, in which uniquely marked golf tees were placed randomly in a field, and records were kept of which tees were observed on each of T sampling occasions. If f_i denotes the number of tees that were found i times, then the resulting data are shown below.

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
46	28	21	13	23	14	6	11

The purpose of the study was to simulate a standard ecological investigation, aimed at estimating animal abundance. For a study of this kind, the likelihood in general consists of components such as:

$$\psi_i = \left\{ \binom{T}{i} p^i (1-p)^{T-i} \right\}^{f_i},$$

where p denotes the probability that any tee is found on any of the sampling occasions.

(a) Explain how you would modify the binomial model in the expression for ψ_i to account for heterogeneity in the probability of finding golf tees, and derive the modified form of ψ_i in terms of beta functions. [12 marks]

Five distinct models have been fitted to the St. Andrews data. Each model is given a name in the table below, and the table presents the value of the maximum log-likelihood for each model, the χ^2 goodness-of-fit statistic, and the number, n_p, of parameters in the model.

Model	-Max log-lik	χ^2	n_p
Bin	104.9	923.2	2
Beta-Bin	22.6	8.4	3
LNB	22.6	8.8	3
2 Bins	27.3	21.8	4
Bin + Beta-Bin	22.2	8.1	5

The number of cells for each of the χ^2 tests was 8.

(b) Construct the AIC values for the 5 models. [8 marks]

(c) Explain the difference between comparing the models in terms of the AIC, and measuring the absolute goodness-of-fit of each model using the χ^2 test. Explain which model or models you believe is/are the best for the St. Andrews data. [10 marks]

Note that in order to answer the last part of the question it is not necessary to understand how each model is constructed.