Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2011–2012**

MAS472 Computational Inference **2 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
*Marks will be awarded for your best **three** answers. Total marks 90.*

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 (i) A random variable X from the Rayleigh distribution has density function

$$f(x) = x \exp\left(-\frac{x^2}{2}\right) \quad x \geq 0.$$

(a) Explain carefully how, given a random number U drawn from a $U[0,1]$ distribution, you can generate a draw from this distribution. ***(6 marks)***

(b) Produce two negatively correlated values from the Rayleigh distribution using this method given a draw $U = 0.71$. ***(2 marks)***

(ii) Suppose instead we wish to generate a random value from the Rayleigh distribution using rejection sampling and a exponential $Exp(1)$ envelope function with density

$$g(y) = \exp(-y) \quad y \geq 0.$$

(a) Show that

$$c = \sup_x \frac{f(x)}{g(x)} = \frac{1+\sqrt{5}}{2} \exp\left\{\frac{-1+\sqrt{5}}{4}\right\} \approx 2.20$$

(5 marks)

(b) Explain carefully how to generate a random value of X given a draw $Y \sim Exp(1)$ and a value $U \sim U[0,1]$. If $Y = 0.4$, given the value $U = 0.6$, determine whether Y is accepted as a random sample from the distribution of X. ***(5 marks)***

(c) Using this envelope function, what is the probability that the first two candidate values are accepted? ***(2 marks)***

(d) An alternative $N(0,1)$ envelope function has been suggested. Would this be feasible and why? ***(2 marks)***

(iii) We also wish to sample from a Weibull$(1,4)$ distribution with density

$$h(x) = 4x^3 \exp\left(-x^4\right) \quad x \geq 0.$$

(a) It is proposed to use importance sampling using a normal approximation. By considering a Taylor series expansion of $\log h(x)$ about the mode of x obtain the mean and variance of the importance density. ***(8 marks)***

2 A sample of independent random observations $X = \{X_1, \ldots, X_n\}$ are drawn from the following mixture distribution:

$$X_i \sim \begin{cases} Poisson(\lambda) & \text{with probability } w \\ Geometric(p) & \text{with probability } 1-w \end{cases}$$

Here the Geometric random variable is defined to count the number k of failures before the first success and so has density

$$P(Z = k) = (1-p)^k p \quad \text{for } k = 0, 1, \ldots$$

with mean $\dfrac{1-p}{p}$. The Poisson has density

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

with mean λ.

Define $\theta = (\lambda, p, w)$ to be the vector of the three unknown parameters. The corresponding 'missing' variables $Y = \{Y_1, \ldots, Y_n\}$ are defined as follows:

$$Y_i = \begin{cases} 1 & \text{if } X_i \text{ is drawn from } Poisson(\lambda), \\ 0 & \text{if } X_i \text{ is drawn from } Geometric(p). \end{cases}$$

(i) Write down the complete data log-likelihood $\log f(X, Y|\theta)$ and use the factorization theorem to show that the sufficient statistics for θ are

$$\begin{aligned} S_1(X,Y) &= \sum X_i Y_i, \\ S_2(X,Y) &= \sum X_i (1 - Y_i), \\ S_3(X,Y) &= \sum Y_i. \end{aligned}$$

(6 marks)

(ii) Derive expressions for $E\{S_1(X,Y)|\theta\}$, $E\{S_2(X,Y)|\theta\}$, $E\{S_3(X,Y)|\theta\}$.
(7 marks)

(iii) Let $\theta_{old} = (\lambda_{old}, p_{old}, w_{old})$ denote an initial estimate θ.

(a) Derive an expression for $h_i = E(Y_i|X, \theta_{old})$. ***(4 marks)***

(b) Derive expressions for $E\{S_1(X,Y)|X, \theta_{old}\}$, $E\{S_2(X,Y)|X, \theta_{old}\}$, $E\{S_3(X,Y)|X, \theta_{old}\}$.
You may leave your expressions in terms of h_i. ***(7 marks)***

(iv) Using your results from parts (ii) and (iii), apply one iteration of the EM algorithm to obtain expressions for an improved estimate of θ. ***(6 marks)***

3 (i) We want to estimate

$$R = \int_{-3}^{2} g(x)dx = \int_{-3}^{2} \exp\left(-\frac{1}{6}|x-1|^{3.5}\right) dx$$

using Monte-Carlo integration. We sample $X_1, \ldots, X_{1000}$ from a $U[-3, 2]$ and it is found that

$$\begin{aligned} \bar{g}(X) = \frac{1}{n}\sum_{i=1}^{1000} g(X_i) &= 0.94 \\ \frac{1}{n-1}\sum_{i=1}^{1000} \{g(X_i) - \bar{g}(X)\}^2 &= 0.037 \end{aligned}$$

(a) Give an approximate 95% confidence interval for R. ***(5 marks)***

(b) Determine how large the Monte Carlo sample size should be such that the width of a 95% confidence interval would be no larger than 0.01. ***(4 marks)***

(c) It is instead suggested to use a alternative $N(0, 1)$ sampling distribution for $X_1, \ldots, X_n$ above. Give the formula for the Monte Carlo estimate of R using this distribution and the new sample $X_1, \ldots, X_n$. ***(4 marks)***

(ii) The monthly maximum values of rainfall are modelled by a Gumbel type-II distribution with density

$$f(x; a, b) = abx^{-(1+a)} \exp\{-bx^{-a}\}$$

for $x > 0$.

(a) Derive the profile log-likelihood function for a given $x_1, \ldots, x_n$. ***(10 marks)***

(b) The rainfall in Zanzibar is measured over the course of a year and the twelve monthly maximum values $X_1, \ldots, X_{12}$ are recorded as follows:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0.92	0.77	0.91	0.92	1.04	0.87	1.01	1.00	1.20	2.63	0.86	0.73

[For the above sample $\sum \log X_i = 0.08$, $\sum X_i^{-5.6} = 22.6$ and $\sum X_i^{-3} = 15.2$.]

The log-likelihood for this data is maximized at $a = 5.6$ and $b = 0.5$. By considering the profile-deviance function, test the null hypothesis that $a = 3$. ***(7 marks)***

4 (i) Within Sheffield there are three climbing walls and there is much argument about where it is best to train. After a joint competition the scores of three random individuals from each of the climbing walls were recorded. The information was entered into `R` in the variables `Wall` and `Score` and is shown below:

```
> Wall
[1] a a a b b b c c c
Levels: a b c
> Score
[1] 162 181 135 179 255 192 205 184 164
```

The following R code is used for analysis:

```
> lm1 <- lm(Score~Wall)
> T <- summary(lm1)$fstatistic[1]
>
> n <- 1000
> K <- rep(NA, n)
> for(i in 1:n) {
+ Samp <- sample(Score, replace = TRUE)
+ lmSamp <- lm(Samp~Wall)
+ K[i] <- summary(lmSamp)$fstatistic[1]
+ }
> sum(K >= T)
[1] 185
```

(a) Explain carefully the procedure that has been used here. State the null hypothesis being tested and the alternative. ***(6 marks)***

(b) Sketch the empirical cumulative distribution function for the scores under the null hypothesis. ***(5 marks)***

(c) What is the outcome of this test? ***(2 marks)***

(d) Comment briefly on the accuracy of this procedure in relation to both the size of n and the number of individuals in the original sample (i.e. 12). ***(3 marks)***

(ii) In another competition, only two climbing walls are present with fewer competitors. The individual scores of those taking part are given by:

Wall a	8, 15, 23
Wall b	21, 27, 29

(a) Conduct an exact randomization test of the null hypothesis of no difference between the two group means, against a one-sided alternative that the mean score from Wall b is higher. Report the level of significance of your observed test statistic. What is the smallest p-value that could be attained with this size of data? ***(12 marks)***

(b) What assumptions must be made to justify this test procedure? ***(2 marks)***

End of Question Paper