Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS Spring Semester 2010–2011

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 An economic model has been constructed to predict the cost per patient of a new treatment for osteoporosis. In addition to the price of the drug, the model includes costs of hospital visits, nursing care, and any necessary surgery. The model has two uncertain inputs:
x: the time in days until the patient's first hip fracture
y: the number of days of nursing home care required, if the patient suffers a fracture.

M is defined to be the expected cost per patient, and P_1 is the probability that a patient's cost will exceed 100,000 pounds.

The following distributions are assumed for x and y:

$$\begin{aligned} x &\sim exponential(rate = 1/30), \\ y &\sim N(20, 25). \end{aligned}$$

The model is implemented in R using a user-defined function `cost(x,y)`. If `x` and `y` are vectors, then `cost(x,y)` will return the appropriate vector output. Some output from the R session is given below.

```
> n<-1000
> x<-rexp(n,1/30)
> y<-rnorm(n,20,5)
> c1<-cost(x,y)
> mean(c1)
[1] 55538.2
> var(c1)
[1] 364736611
> sum(c1>100000)
[1] 72
```

(i) Estimate M and P_1, giving 95% confidence intervals for both. What would you expect to happen to your 95% confidence interval for P_1 if `n` were changed to 100000? ***(10 marks)***

1 (continued)

(ii) A modification to the listing is made:

```
> u.1<-runif(500,0,1)
> u.2<-1-u.1
> x<-c( -30*log(1-u1) , -30*log(1-u2) )
> y<-rnorm(n,20,5)
> c1<-cost(x,y)
> mean(c1)
[1] 60483.1
> var(c1)
[1]  341901123
> cor(c1[1:500],c1[501:1000])
[1] -0.511
```

(a) Regarding the first three lines, name the two techniques that have been used to produce `x`. Give the motivation for this modification to the code. ***(3 marks)***

(b) Based on the new output, calculate a 95% confidence interval for M. ***(9 marks)***

(iii) An alternative distribution is proposed for y: the $Gamma(10, 0.5)$ distribution, with density function

$$f(y) = \frac{0.5^{10}}{\Gamma(10)} y^9 \exp(-0.5y),$$

for $y > 0$.

(a) If the original R analysis generated output values $c_1, \ldots, c_{1000}$ from input values $x_1, \ldots, x_{1000}$ and $y_1, \ldots, y_{1000}$, give a formula for the Monte Carlo estimate of M, corresponding to the new distribution of y, in terms of $c_1, \ldots, c_{1000}$ and $y_1, \ldots, y_{1000}$, which could be calculated without doing any further evaluations of the function `cost`. ***(5 marks)***

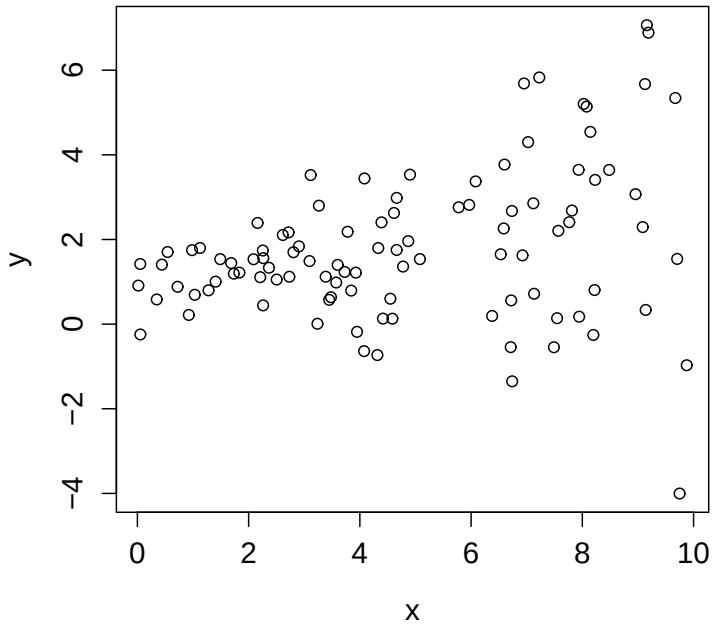
(b) Explain how you would calculate a 95% confidence interval for M in this case.

(3 marks)

2 Given observations (x_i, y_i) for $i = 1, \ldots, 100$, the following model is considered:

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, 100$. The data are plotted below.



In an R session, with $(y_1, \ldots, y_{100})$ and $(x_1, \ldots, x_{100})$ stored under the variable names `y` and `x` respectively, the least squares estimate of β is obtained with the commands

```
lm1<-lm(y~x)
lm1$coefficients[2]
```

The least squares estimator $\hat{\beta}$ of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{100}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{100}(x_i - \bar{x})^2}.$$

2 (continued)

(i) Below is some output from an R session.

```
k<-rep(0,1000)
for(i in 1:1000){
z<-sample(x,size=100,replace=F)
lm2<-lm(y~z)
k[i]<-lm2$coefficients[2] }

lm1<-lm(y~x)
a<-lm1$coefficients[2]
sum(abs(k)<abs(a))
[1] 998
```

(a) State the null hypothesis being tested, explain the procedure that has been used to conduct the test, and give the result of the test. Why might this procedure be preferable to a test derived from the assumption that $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, 100$? ***(9 marks)***

(b) Derive a simpler function of the data that could be used in definitions of `k[i]` and `a`, without changing the result (assuming the same seed was used in each case). ***(7 marks)***

(c) If there were only 5 observations rather than 100, what is the smallest possible p-value that could be obtained using this procedure? ***(5 marks)***

(ii) Some further R code from the same session is given below.

```
k<-rep(0,1000)
for(i in 1:1000){
a<-sample(c(1:100),size=100,replace=T)
r<-y[a]
z<-x[a]
lm2<-lm(r~z)
k[i]<-lm2$coefficients[2] }

quantile(k,c(0.05,0.95))
        5%        95%
0.06825594 0.31096000
```

Give the name and a brief description of the statistical procedure that has been used here. What do the two numbers in the last line represent? ***(4 marks)***

2 (continued)

(iii) Given observations (x_i, y_i) for $i = 1, \ldots, 100$, consider the model

$$y_i = \beta x_i + \varepsilon_i,$$

with $\varepsilon_i \sim t_1$, for $i = 1, \ldots, 100$, so that each error has a Student-t distribution with 1 degree of freedom. Explain carefully the steps required to perform a Monte Carlo test of size 0.01 of the hypothesis $\beta = 0$, using the least squares estimator of β as a test statistic. You are not required to provide any R code. ***(5 marks)***

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

$$X_i \sim \begin{cases} N(\mu, \sigma^2) & \text{with probability } \phi \\ Exponential(\text{rate} = \lambda) & \text{with probability } 1 - \phi \end{cases}$$

Define $\theta = (\mu, \sigma^2, \lambda, \phi)$ to be the vector of the four unknown parameters. The corresponding 'missing' variables $Y = \{Y_1, \ldots, Y_n\}$ are defined as follows:

$$Y_i = \begin{cases} 1 & \text{if } X_i \text{ is drawn from } N(\mu, \sigma^2), \\ 0 & \text{if } X_i \text{ is drawn from } Exponential(\text{rate}= \lambda). \end{cases}$$

However, X_1 and X_2 are observed to be negative, taking distinct values, and so $Y_1 = 1$ and $Y_2 = 1$ are known.
Sufficient statistics for $\theta = (\mu, \sigma^2, \lambda, \phi)$ are

$$\begin{aligned} S_1(X, Y) &= \sum_{i=1}^{n} Y_i, \\ S_2(X, Y) &= \sum_{i=1}^{n} Y_i X_i, \\ S_3(X, Y) &= \sum_{i=1}^{n} Y_i X_i^2, \\ S_4(X, Y) &= \sum_{i=1}^{n} X_i \end{aligned}$$

(i) Write down the complete data log-likelihood function for $\theta = (\mu, \sigma^2, \lambda, \phi)$ given both X and Y. ***(5 marks)***

(ii) Derive expressions for $E\{S_1(X, Y)|\theta\}$, $E\{S_2(X, Y)|\theta\}$, $E\{S_3(X, Y)|\theta\}$ and $E\{S_4(X, Y)|\theta\}$. ***(8 marks)***

(iii) Let $\theta_{old} = (\mu_{old}, \lambda_{old}, \sigma^2_{old}, \phi_{old})$ denote an initial estimate θ.

(a) Derive an expression for $p_i = E(Y_i|X, \theta_{old}, Y_1, Y_2)$, for $i > 2$. ***(5 marks)***

(b) Derive expressions for $E\{S_1(X, Y)|X, Y_1, Y_2, \theta_{old}\}$, $E\{S_2(X, Y)|X, Y_1, Y_2, \theta_{old}\}$, $E\{S_3(X, Y)|X, Y_1, Y_2, \theta_{old}\}$ and $E\{S_4(X, Y)|X, Y_1, Y_2, \theta_{old}\}$. You may leave your expressions in terms of p_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. ***(5 marks)***

4 (i) Random variables $X_1, \ldots, X_n$ are independent and each have the inverse gamma (a, b) distribution, with density function given by

$$f(x) = \frac{b^a x^{-(a+1)} \exp\left(-\frac{b}{x}\right)}{\Gamma(a)},$$

for $x > 0$, and for integer n, $\Gamma(n) = (n-1)!$

(a) Derive the profile log-likelihood function for a. ***(9 marks)***

(b) For a sample of size 15, it is observed that $\sum_{i=1}^{15} x_i^{-1} = 20.983$ and $\sum_{i=1}^{20} \log x_i = -2.731$. If the profile log-likelihood function for a is maximised at $\hat{a} = 3.412$, give the maximum likelihood estimate of b, and use the profile deviance function to test the null hypothesis $H_0 : a = 2$. Note that $\Gamma(3.412) = 3.0198$. ***(7 marks)***

(ii) Suppose it is desired to sample from the t_3 distribution: the Student-t distribution with 3 degrees of freedom. The density function is given by

$$f(x) = \frac{4\sqrt{3}}{\pi}\left(1 + \frac{x^2}{3}\right)^{-2}.$$

Importance sampling is to be used, with an importance density based on approximating $f(x)$ by a normal density function.

(a) By considering a Taylor series expansion of $\log f(x)$ about 0, obtain the mean and variance of the importance density. ***(8 marks)***

(b) Given two random draws $U_1 = 0.90$ and $U_2 = 0.45$ from the $U[0, 1]$ distribution, obtain two samples from the t_3 distribution via the importance density, and calculate the weights of your two sampled values. ***(6 marks)***

End of Question Paper