Queen Mary
University of London

M. Sci. Examination by course unit 2014

MTH731U: Computational Statistics

Duration: 3 hours

Date and time: 9 May 2014, 10:00h–13:00h

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

You should attempt all questions. Marks awarded are shown next to the questions.

Calculators ARE permitted in this examination. The unauthorised use of material stored in pre-programmable memory constitutes an examination offence. Please state on your answer book the name and type of machine used.

Statistical functions provided by the calculator may be used provided that you state clearly where you have used them.

The New Cambridge Statistical Tables are provided.

Complete all rough workings in the answer book and cross through any work which is not to be assessed.

Important note: the Academic Regulations state that possession of unauthorised material at any time by a student who is under examination conditions is an assessment offence and can lead to expulsion from QMUL.

Please check now to ensure you do not have any notes, mobile phones or unauthorised electronic devices on your person. If you have any, then please raise your hand and give them to an invigilator immediately. Please be aware that if you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. Disruption caused by mobile phones is also an examination offence.

Exam papers must not be removed from the examination room.

Examiner(s): H. Grossmann and D. S. Coad

Question 1 (22 marks)

(a) State the general form of a kernel estimator of a probability density function f explaining all terms. Which component of a kernel estimator has the strongest influence on the appearance of the estimated probability density function? [4]

(b) Which of the two functions

$$K_1(x) = \begin{cases} \frac{\pi}{4}\sin(\frac{\pi}{2}x) & -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$K_2(x) = \begin{cases} \frac{\pi}{4}\cos(\frac{\pi}{2}x) & -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

could be used as a kernel function? Justify your answer. [6]

(c) For the data

$$-0.5, \qquad -0.3, \qquad 0.1, \qquad 0.3$$

compute Rosenblatt's estimator with bandwidth $h = 0.6$ and sketch its graph. [8]

(d) Suppose that in part (c) the R code

```
y <- c(-0.5, -0.3, 0.1, 0.3)
d <- density(y, bw=0.6, kernel="rectangular")
```

was used to find Rosenblatt's estimator.

Would this give the correct result? Justify your answer. [4]

Question 2 (21 marks) Programming languages usually provide a random number generator for producing numbers which are claimed to be independent draws from the uniform distribution on the open interval $(0, 1)$. The following ten values, which have been rounded to four decimal places, were produced by ten runs of such a random number generator:

$$\begin{array}{ccccc} 0.2213, & 0.3528, & 0.5852, & 0.5612, & 0.1395, \\ 0.6345, & 0.6168, & 0.7771, & 0.2238, & 0.7550. \end{array}$$

(a) Test at the 10% significance level the hypothesis that the data are a sample from the uniform distribution on $(0, 1)$. State the hypotheses for the test and your conclusion. [8]

(b) Explain why using for the test in part (a) a 10% significance level may be more appropriate than using, for example, a 1% level of significance. [3]

(c) Let Y denote the random variable which represents the result of a single run of the random number generator. Suppose that Y is uniformly distributed on the interval $(0, 1)$.

Obtain the distribution of the random variable $X = -2\log(1 - Y)$. [6]

(d) Briefly explain how by using part (c) one can generate random numbers which have an exponential distribution with parameter $\lambda = 1/2$. [4]

Question 3 (24 marks)

(a) Consider observations $y_1, \dots, y_n$ of independent and identically distributed random variables $Y_1, \dots, Y_n$ with cumulative distribution function F. Let θ be a parameter of the distribution represented by F. For an estimator $\hat{\theta}$ of θ, define the jackknife estimators of the bias and of the variance of $\hat{\theta}$, explaining all terms. [6]

(b) Salaries in a population of interest are often modelled by the Pareto distribution with probability density function

$$f(y) = \theta k^{\theta} y^{-(\theta+1)} \quad \text{for } y \geq k,$$

where $k > 0$ represents some minimum salary and $\theta > 1$. Economists are particularly interested in estimating the parameter θ from a sample $y_1, \dots, y_n$. One estimator of θ is

$$\hat{\theta} = \frac{n\bar{y} - y_{(1)}}{n(\bar{y} - y_{(1)})}, \tag{1}$$

where $\bar{y}$ is the sample mean and $y_{(1)}$ is the smallest salary in the sample.

(i) Suppose the jackknife is used to estimate the variance of $\hat{\theta}$ from $y_1, \dots, y_n$. During the calculation of the jackknife estimate of variance, will $y_{(1)}$ in equation (1) always have the same value? Justify your answer. [3]

(ii) The following data are annual salaries in units of £1,000 per year:

43, 24, 60, 55, 27.

Calculate the jackknife estimates of the bias and of the variance of $\hat{\theta}$. [7]

(c) Prove that for a sample $y_1, \dots, y_n$ the jackknife estimate of variance of the sample mean $\bar{y}$ is equal to $\widehat{var}_{\text{jack}} = \frac{1}{n}s^2$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance. [8]

Question 4 (10 marks) Consider the simple linear regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where Y_i is the random variable representing the response at the value x_i of the explanatory variable and the ϵ_is are uncorrelated random errors with zero means and equal variances σ^2. If the assumptions about the ϵ_is are in doubt, a bootstrap approach may be considered.

Give a step-by-step description of how the method of *bootstrapping cases* would be applied to a sample $(x_1, y_1), \dots, (x_n, y_n)$ in order to estimate the bias and the standard error of the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ of the intercept α and the slope β. [10]

Question 5 (9 marks) Consider the following lines of R code:

```
u <- c(-1.3,0.5,2.1,2.3,-0.6,1.1,-0.4,2.8,1.8,0.2)
v <- ppoints(length(u), a=0.5)
w <- qnorm(v)
plot(w, sort(u))
```

(a) What outcome does this code produce? [3]

(b) Which values will be contained in `v` after the second line has been executed? [3]

(c) Explain the meaning of the command `w <- qnorm(v)`. [3]

Question 6 (14 marks) Three patients suffering from chronic pain were trained to use coping strategies for three weeks. Pain scores were recorded before and after the training. The pain scores are shown below, where smaller values correspond to less pain:

	Patient		
	1	2	3
Before	5.3	4.7	3.0
After	1.8	2.7	3.5

(a) Compute the p-value of a suitable nonparametric test for testing if the training was able to reduce the pain. Use the p-value to test the hypothesis at the 5% level of significance. [8]

(b) List two other types of experimental situations or studies where the test you have used in (a) is appropriate. [3]

(c) Briefly explain the main difference between the parametric and nonparametric approaches to hypothesis testing. [3]

End of Paper