Queen Mary
University of London

Main Examination period 2017

MTH731U / MTHM731 / MTH731P: Computational Statistics

Duration: 3 hours

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 available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.
The New Cambridge Statistical Tables are provided.

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

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it shall 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. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: J. Griffin, L. Pettit

 Turn Over

Question 1. [10 marks]

(a) Suppose that we want to graphically check if a sample is consistent with some continuous probability distribution, called the reference distribution. One way of doing this is a Q-Q plot. Explain what pair of values each plotted point represents in this type of graph. If the sample is from the reference distribution, what general pattern would we expect to see? [**4**]

(b) Assume that the reference distribution is a standard normal distribution. Draw a sketch of how the Q-Q plot would appear if the sample was from a normal distribution with a mean of 10 and standard deviation 5. Also draw a sketch of the Q-Q plot we would see if the sample was from an exponential distribution with mean 1. [**6**]

Question 2. [13 marks]

(a) Suppose we have a random sample $y_1, \dots, y_n$. Define the empirical cumulative distribution function (ecdf) for this sample. [**3**]

(b) The Kolmogorov-Smirnov statistic is given by

$$D_n = \max\left(D_n^+, D_n^-\right)$$

where

$$D_n^+ = \sup_{y \in \mathbb{R}}[\hat{F}_n(y) - F_0(y)] \quad \text{and} \quad D_n^- = \sup_{y \in \mathbb{R}}[F_0(y) - \hat{F}_n(y)].$$

with $\hat{F}_n$ being the ecdf and F_0 a continuous cumulative distribution function that we are testing for agreement with. Let $y_{(1)}, \dots, y_{(n)}$ be the ordered values of the random sample. Note that we can also write

$$D_n^+ = \max_{y \in \mathbb{R}}[\hat{F}_n(y) - F_0(y)] \quad .$$

(i) When calculating D_n^+, explain why for each interval $y_{(i)} \le y < y_{(i+1)}$, with $i < n$, we only need to consider the value of $\hat{F}_n(y) - F_0(y)$ at $y = y_{(i)}$ and not on the rest of the interval. [**4**]

(ii) Based on the idea in part (b)(i), prove that

$$D_n^+ = \max_{1 \le i \le n}[\hat{F}_n(y_{(i)}) - F_0(y_{(i)})]$$

[**6**]

Question 3. [12 marks]
Let $x_1,\dots,x_m$ and $y_1,\dots,y_n$ be two independent random samples, and suppose that all $m+n$ values are distinct.

(a) Define the Mann-Whitney statistic U_X for these samples based on the ranks of $x_1,\dots,x_m$. **[4]**

(b) Show that if both samples are generated by the same continuous probability distribution, then
$$E(U_X) = \frac{mn}{2}.$$
[8]

Question 4. [14 marks]

(a) Pain scores were obtained for three patients before and after receiving medication.

Patient	1	2	3
After	1.87	1.71	1.73
Before	2.64	1.84	2.31

We want to find out if the treatment has led to a decrease in the pain scores without making a normality assumption. Use an appropriate permutation test to test this hypothesis at the 10% level of significance. In your answer, calculate the full null distribution. **[10]**

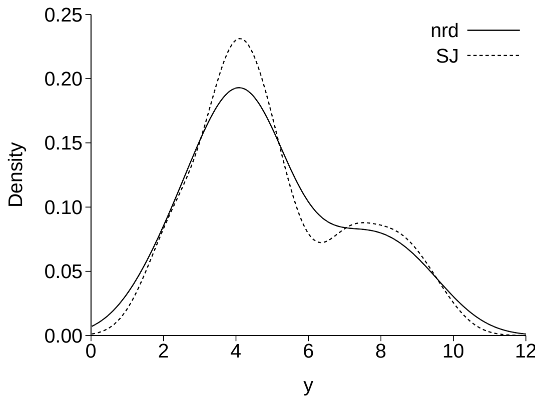
(b) Suppose that in part (a), we wanted to carry out the test at the 1% significance level. What is the minimum number of patients we would need in order for it to be possible for us to reject the null hypothesis? **[4]**

Question 5. [16 marks]

(a) State the general formula for a kernel density estimator (KDE) of a probability density function f explaining all terms. **[4]**

(b) For a given sample size, how do the bias and variance of a KDE at a single point change as the bandwidth is made smaller? **[4]**

(c) The plot below shows two kernel density estimates for the same sample using two methods for finding the bandwidth, which in the R command "density" are referred to as "nrd" and "SJ".



The optimal bandwidth h_n that minimises the asymptotic mean squared error is given by

$$h_n = \left(\frac{A}{n\sigma_K^4 \int_{-\infty}^{+\infty} (f''(y))^2 dy} \right)^{\frac{1}{5}}$$

where A and σ_K are constants, n is the sample size and f is the unknown density function.

(i) Explain briefly, without going into mathematical details, how the method "nrd" uses the formula for h_n. **[3]**

(ii) What could cause the difference in appearance between the two estimates that are plotted, and why would the method "SJ" lead to this difference? **[5]**

Question 6. [12 marks]
Suppose that we have bivariate data of the form $(y_1, x_1), \ldots, (y_n, x_n)$. We wish to fit models of the form $E(Y_i) = f(x_i, \boldsymbol{\beta})$, where f is a known functional form and $\boldsymbol{\beta}$ is a vector of parameters to be estimated.

(a) Describe the procedure for using leave-one-out cross-validation to obtain a set of predictions $\hat{y}_{[1]}, \ldots, \hat{y}_{[n]}$. **[4]**

(b) Define the predicted residuals that result from the leave-one-out cross-validation procedure. If we are fitting a linear model, how do the predicted residuals compare in magnitude to the ordinary residuals that we get when fitting the model to the original dataset? **[4]**

(c) Define the PRESS statistic and explain how we can use it to choose among several possible models with different forms for f and $\boldsymbol{\beta}$. **[4]**

Question 7. [23 marks]

(a) If we have a dataset of distinct values $y_1 \ldots, y_n$, state briefly how we would generate a set of leave-one-out jackknife replications for some estimator $\hat{\theta}$. If $\hat{\theta}$ is the sample median and $n = 100$, how many different values will the jackknife replications take? If instead $\hat{\theta}$ is the sample mean and $n = 100$, how many different values will the jackknife replications take? [**8**]

(b) Consider the simple linear regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

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 $(x_1, y_1), \ldots, (x_n, y_n)$ in order to estimate the standard error of the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ of the intercept α and the slope β. [**9**]

(c) Explain how the procedure in part (b) would be modified if we instead want to bootstrap residuals. [**6**]

End of Paper.