Queen Mary
University of London

Main Examination period 2018

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.

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, H. Maruri-Aguilar

Question 1. [9 marks]
Consider the following lines of R code:

```
v = c(1.4,0.7,0.2,1.5,-1.9,2.2,-0.8,1.3,1.1,0.6)
n = length(v)
p = ((1:n)-0.5)/n
q = qnorm(p)
plot(q, sort(v))
```

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

(b) What values will `p` contain after the third line has been executed? [**3**]

(c) Explain the meaning of the command `q = qnorm(p)`. [**3**]

Question 2. [21 marks]

(a) Define the Wilcoxon signed-rank statistic W^+ for a sample $z_1, \dots, z_n$. [**3**]

(b) What would be the null hypothesis for the Wilcoxon signed-rank test? Show that under this null hypothesis for a sample of size n, W^+ has a mean

$$E(W^+) = \frac{n(n+1)}{4}.$$

[**8**]

(c) Blood cholesterol measurements were taken for eight patients before and after a course of medication.

Patient	1	2	3	4	5	6	7	8
Before	5.7	6.2	7.4	6.0	5.1	7.3	6.9	6.4
After	6.5	4.1	5.0	6.3	4.2	3.3	5.2	3.8

We want to find out if the medication has led to a decrease in the cholesterol level. Use an appropriate rank test to test this hypothesis at the 5% level of significance. [**10**]

Question 3. [15 marks]

(a) Consider the samples $6.8, 5.3, 7.1$ and $4.2, 5.9$ from two populations. We want to know if the mean in the population associated with the first sample is different from the population mean for the second sample. We are not prepared, however, to assume that the data are normally distributed.

Suppose we want to perform a permutation test. State an appropriate null hypothesis and a test statistic. Perform a permutation test at the 10% significance level to test the hypothesis. [**12**]

(b) For sample sizes which are too large to calculate the exact null distribution, even by computer, explain how we might approximate the null distribution for a permutation test. [**3**]

Question 4. [20 marks]

(a) Consider the following data:

$$1.7,\ 5.9,\ 7.2,\ 2.8,\ 5.1,\ 5.6,\ 6.4,\ 4.4,\ 2.5,\ 4.6$$

Find the histogram estimator of the probability density function $\hat{f}_H(y)$ for all $y \in \mathbb{R}$, taking the interval end points to be $0, 2, 4, 6, 8$. [**8**]

(b) For a given sample size, how do the bias and variance of the histogram estimator $\hat{f}_H(y)$ at a single point y change as the interval width decreases? [**4**]

(c) State the general formula for a kernel density estimator (KDE) of a probability density function, explaining all terms. Which component of a KDE has the strongest influence on the appearance of the estimated density? [**5**]

(d) State one advantage of using a KDE rather than a histogram for estimating a probability density. [**3**]

Question 5. [16 marks]

(a) Suppose that we have two random samples $\mathbf{x} = x_1, \dots, x_m$ and $\mathbf{y} = y_1, \dots, y_n$, which are assumed to be independent of each other. Let θ be the quantity we are interested in, the ratio of the standard deviation of the first population and the the standard deviation of the second population, which we estimate using

$$\hat{\theta} = \sqrt{\frac{Var(\mathbf{x})}{Var(\mathbf{y})}}$$

Describe how we would generate a non-parametric bootstrap sample for $\hat{\theta}$, and how we would use this sample to estimate the standard error of $\hat{\theta}$. [**9**]

(b) Explain how we would calculate a 95% percentile confidence interval for $\hat{\theta}$ using the bootstrap sample in (a). [**4**]

(c) In general, if a confidence interval (θ_L, θ_U) is calculated for a population parameter θ, define what is meant by the *coverage* of the confidence interval. [**3**]

Question 6. [19 marks]
Suppose that our data is of the form $(y_1, x_1), \ldots, (y_n, x_n)$. We wish to fit models of the form $E(Y) = g(x;\beta)$, where β is a vector of parameters to be estimated.

(a) Describe the general procedure for using leave-one-out cross-validation to obtain a set of predictions $\hat{y}_{[1]}, \ldots, \hat{y}_{[n]}$. [**5**]

(b) Define the predicted residuals that result from the leave-one-out cross-validation procedure, and define the PRESS statistic. [**4**]

(c) Suppose that g depends on a set of spline functions and we estimate β by minimizing the penalized sum of squares

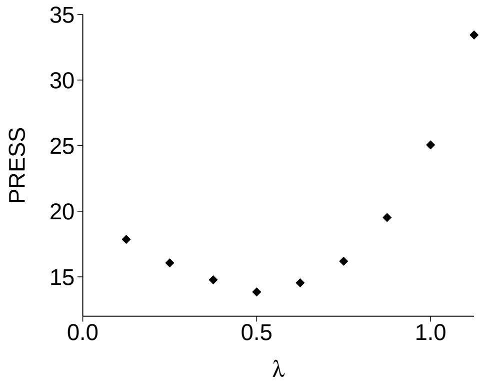
$$PSS = \sum_{i=1}^{n} (y_i - g(x_i;\beta))^2 + \lambda \int_{-\infty}^{\infty} g''(x;\beta)^2 \, dx$$

where $\lambda > 0$ is a smoothing parameter.

The answers do not need any details about spline functions.

(i) Explain why for sufficiently large values of λ, the fitted model approaches a linear function. [**3**]

(ii) Suppose that we have fitted this model for a range of values of λ and calculated the PRESS statistic each time, with the results as plotted below.



Explain how this graph would be used to select a value of λ. By doing this, what feature of the fitted model are we selecting for? Why would PRESS initially decrease as λ increases for small values of λ? [**7**]

End of Paper – An appendix of 2 pages follows.

Appendix: statistical tables

Normal distribution function

Table 1: The standard normal cumulative distribution function $\Phi(x)$ for the given values of x. The cdf for $x < 0$ can be found using the fact that $\Phi(x) = 1 - \Phi(-x)$. For $x \geq 3.8$, $1 - \Phi(x) < 10^{-4}$.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.0	0.500	0.8	0.788	1.6	0.945	2.4	0.992	3.2	0.9993
0.1	0.540	0.9	0.816	1.7	0.955	2.5	0.994	3.3	0.9995
0.2	0.579	1.0	0.841	1.8	0.964	2.6	0.995	3.4	0.9997
0.3	0.618	1.1	0.864	1.9	0.971	2.7	0.997	3.5	0.9998
0.4	0.655	1.2	0.885	2.0	0.977	2.8	0.997	3.6	0.9998
0.5	0.691	1.3	0.903	2.1	0.982	2.9	0.998	3.7	0.9999
0.6	0.726	1.4	0.919	2.2	0.986	3.0	0.999	3.8	0.9999
0.7	0.758	1.5	0.933	2.3	0.989	3.1	0.999		

Table 2: Selected upper quantiles of the standard normal distribution. For each p in the first row, the second row contains the value of x such that $\Phi(x) = 1 - p$.

p	0.1	0.05	0.025	0.01	5×10^{-3}	10^{-3}	5×10^{-4}	10^{-4}
$\Phi^{-1}(1-p)$	1.28	1.64	1.96	2.33	2.58	3.09	3.29	3.72

Wilcoxon signed-rank critical values

Table 3: Lower critical values of the one-sample Wilcoxon signed-rank statistic W^+ for samples of size n. For each n and P, the entry in the table is the largest value x such that $P(W^+ \leq x) \leq P$. If there is no such x, then the entry is blank (-).

P n	0.05	0.025	0.01	0.005	0.001
5	0	-	-	-	-
6	2	0	-	-	-
7	3	2	0	-	-
8	5	3	1	0	-
9	8	5	3	1	-
10	10	8	5	3	0
11	13	10	7	5	1
12	17	13	9	7	2

Mann-Whitney critical values

Table 4: Lower critical values of the two-sample Mann-Whitney statistic U_X for samples of size m and n under the null hypothesis that both samples have the same distribution. For each m, n and P, the entry in the table is the largest value x such that $P(U_X \leq x) \leq P$. If there is no such x, then the entry is blank (-).

	P	0.05	0.025	0.01	0.005	0.001
m	n					
2	5	0	-	-	-	-
2	6	0	-	-	-	-
2	7	0	-	-	-	-
2	8	1	0	-	-	-
3	3	0	-	-	-	-
3	4	0	-	-	-	-
3	5	1	0	-	-	-
3	6	2	1	-	-	-
3	7	2	1	0	-	-
3	8	3	2	0	-	-
4	4	1	0	-	-	
4	5	2	1	0	-	-
4	6	3	2	1	0	-
4	7	4	3	1	0	-
4	8	5	4	2	1	-
5	5	4	2	1	0	-
5	6	5	3	2	1	-
5	7	6	5	3	1	-
5	8	8	6	4	2	0
6	6	7	5	3	2	-
6	7	8	6	4	3	0
6	8	10	8	6	4	1
7	7	11	8	6	4	1
7	8	13	10	7	6	2
8	8	15	13	9	7	4

End of Appendix.