Queen Mary
University of London

MTH6931: Computational Statistics

Duration: 2 hours

Date and time: 3rd June 2016, 14:30–16:30

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. The New Cambridge Statistical Tables are provided.

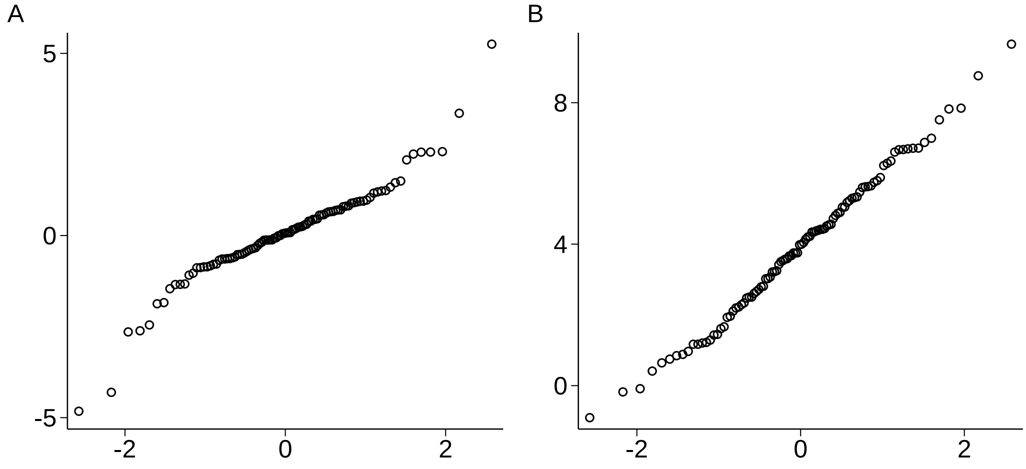
Complete all rough workings in the answer book and cross through any work that is not to be assessed.

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.

Examiner(s): J. Griffin, H. Maruri-Aguilar

Question 1 (16 marks).



(a) Plots A and B show Q-Q plots for two samples of size 100, of different quantities. Assuming that the horizontal axes show the reference distribution, explain for this type of plot what is shown on the axes, what each circle in the plots represents and what general pattern we expect to see if the sample is from the reference distribution. **[7]**

(b) Assuming that the reference distribution is a standard normal distribution, what does each plot tell us about the distribution of that sample? **[9]**

Question 2 (23 marks).

(a) Suppose we have a random sample $y_1, \ldots, y_n$. Define the empirical cumulative distribution function for this sample. **[5]**

(b) We wish to test at the 10% level of significance if the sample

$$2.85,\ 3.77,\ 4.98,\ 3.36,\ 5.87$$

comes from a normal distribution with mean 4 and standard deviation 1, using the two-sided Kolmogorov-Smirnov one sample test. Carry out this test, stating clearly your null hypothesis and conclusions. **[18]**

Question 3 (20 marks).

(a) State the general formula for a kernel density estimator (KDE) of a probability density function f explaining all terms. Which component of a KDE has the strongest influence on the appearance of the estimated probability density function? **[6]**

(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? **[6]**

(c) For a KDE with kernel K, where K has variance σ_K^2, the asymptotic mean integrated square error AMISE is for large sample size n and small bandwidth h given by

$$\text{AMISE} = \frac{1}{nh}A + \frac{1}{4}h^4 B$$

where

$$A = \int_{-\infty}^{+\infty} K^2(y)\,dy\,, \quad B = \sigma_K^4 \int_{-\infty}^{+\infty} (f''(y))^2\,dy$$

Find the value h^* of h that minimizes the AMISE, showing that it is a minimum. **[8]**

Question 4 (27 marks).

(a) Consider the samples $5.92, 4.66, 6.30$ and $5.21, 3.32$ from two populations. We wish to know if the mean in the population associated with the first sample is greater than 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 one-sided permutation test at the 10% significance level to test the hypothesis. **[18]**

(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. **[4]**

(c) Briefly explain the main difference between the parametric and nonparametric approaches to hypothesis testing. **[5]**

Question 5 (14 marks).

(a) Data on the distribution of incomes are usually positively skewed. Hence the median is often regarded as being a more appropriate summary than the mean. Give a brief description of how the nonparametric bootstrap can be applied to estimate the standard error when the median is used to estimate typical income from a sample $y_1, \ldots, y_n$ of income data. In the description list the necessary steps and give the definition of the bootstrap estimate of the standard error. [**9**]

(b) A distribution that is often used to model income distributions is the lognormal distribution with probability density function

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\log(y)-\mu)^2}{2\sigma^2}}$$

where μ and σ are the parameters of the distribution. Briefly explain how the procedure in part (a) has to be modified if the parametric bootstrap is to be applied to data which are assumed to have a lognormal distribution. [**5**]

End of Paper.