The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2017–2018**

MAS472 Computational Inference **2 hours**

Candidates may bring to the examination a calculator that conforms to University regulations.
Answer all questions. Total marks 60.

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 (i) Define the cumulative distribution function (CDF), $F_X(x)$, of a random variable X.

(1 mark)

(ii) Suppose you are given a set of independent identically distributed samples from $F_X(\cdot)$, i.e., you are given data $\{X_1, \ldots, X_n\}$ where each X_i has the same distribution as X. Give the mathematical expression for an unbiased estimator of the CDF of X based on these data, $\widehat{F}_X(\cdot)$ say, and prove that it is an unbiased estimator of F.

(5 marks)

(iii) State the asymptotic distribution of $\widehat{F}_X(x)$ (i.e., the distribution as $n \to \infty$).

(3 marks)

(iv) The median, $m(F)$, of distribution F is defined to be the value m such that

$$\int_{-\infty}^{m} \mathrm{d}F(x) \geq \frac{1}{2} \text{ and } \int_{m}^{\infty} \mathrm{d}F(x) \geq \frac{1}{2}.$$

Using the 'plug-in principle', find an estimator of the median of F, i.e., calculate $m(\widehat{F})$.

You may assume that n, the number of samples in your dataset, is an odd number.

(4 marks)

(v) Describe a bootstrap procedure for estimating the standard error of $m(\widehat{F}_X)$.

(5 marks)

(vi) State how you would calculate a 95% confidence interval for $m(\widehat{F}_X)$.

(2 marks)

2 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 £30,000.

The following distributions are assumed for x and y:

$$\begin{aligned} x &\sim exponential(rate = 1/20), \\ y &\sim N(10, 4). \end{aligned}$$

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,1/20)
> y<-rnorm(n,10,2)
> c1<-cost(x,y)
> mean(c1)
[1] 10419.72
> var(c1)
[1] 141763122
> sum(c1>30000)
[1] 65
```

(i) (a) Estimate M and P_1, giving 95% confidence intervals for both.

(6 marks)

(b) How large would n need to be to find a 95% confidence for M that has a width of less than 10?

(1 mark)

2 (continued)

(ii) The R code is now changed as follows:

```
> u1<-runif(500,0,1)
> u2<-1-u1
> x<-c( -20*log(1-u1) , -20*log(1-u2) )
> y<-rnorm(1000,10,2)
> c1<-cost(x,y)
> mean(c1)
[1] 10793.6
> var(c1)
[1] 153930901
> cor(c1[1:500],c1[501:1000])
[1] -0.5053812
```

(a) What two techniques have been used in the first three lines of this code? Give the motivation for these changes to the code. ***(3 marks)***

(b) Based on the new output, calculate a 95% confidence interval for M. ***(5 marks)***

(iii) An alternative distribution is proposed for y: the χ^2_4 distribution, with density function

$$\pi_Y(y) = \frac{1}{4} y e^{-\frac{y}{2}},$$

for $y > 0$.

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)***

3 (i) In the R code below, the vector `x` consists of measurements from 6 patients in a treatment group, and the vector `y` consists of measurements from 6 patients in a control group. Two different methods are used for testing the same hypothesis about the two group means.

Method I:

```
T.obs <- t.test(x,y)$statistic
data <- c(x,y)
T.rand <- c()
for(i in 1:1000){
  d <- sample(data, replace=F)
  T.rand[i] <- t.test(d[1:6],d[7:12])$statistic
}
sum(abs(T.rand) >= abs(T.obs))
```

Method II:

```
smp.mean <- mean(data)
smp.sd <- sd(data)
T.sim <- replicate(10^4,
  {
   x <- rnorm(6, mean=smp.mean, sd=smp.sd)
   y <- rnorm(6, mean=smp.mean, sd=smp.sd)
   t.test(x,y)$statistic
  })
sum(abs(T.sim)>= abs(T.obs))
```

(a) State the null hypothesis being tested and the alternative hypothesis. For each method, name the technique that is being used to implement the hypothesis test. ***(3 marks)***

(b) State the main assumption that is required for method I to be appropriate. ***(1 mark)***

(c) Suppose an exact randomisation test is used to test the same null hypothesis (with the same alternative hypothesis). What is the smallest possible p-value that could be obtained? ***(2 marks)***

(d) If method I gave an output of 41, and method II gave an output of 310, state the results of the two hypothesis tests. ***(1 mark)***

3 (continued)

(ii) Let X have a $N(0,1)$ distribution, truncated to lie inside the interval $[-k,k]$. It has probability density function

$$f(x) = \begin{cases} \frac{r}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} & \text{if } x \in [-k,k] \\ 0 & \text{otherwise,} \end{cases}$$

where r is a constant to be determined.

(a) Derive an expression for r, involving the CDF for the $N(0,1)$ distribution $\Phi(\cdot)$, for a given value of k. ***(3 marks)***

(b) Random values of X can be generated using rejection sampling, using a $U[-k,k]$ distribution as the proposal/envelope function. Derive the rejection sampling algorithm, and state its acceptance rate. ***(6 marks)***

(c) A different rejection algorithm is to generate Y from a $N(0,1)$ distribution, and accept it if $Y \in [-k,k]$, otherwise reject.

Calculate the acceptance rate of this algorithm, and hence find the largest value of k for which using a $U[-k,k]$ proposal has an acceptance rate that is higher than using a standard normal proposal. ***(4 marks)***

End of Question Paper