The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2016–2017**

MAS472 Computational Inference **2 hours**

Candidates may bring to the examination a calculator that conforms to University regulations.
*Marks will be awarded for your best **three** answers. 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 Suppose we are given an iid sample of data $\{X_1, \ldots, X_n\}$ from a distribution with cumulative density function (CDF) $F(\cdot)$, mean $\mu = \mathbb{E}_F(X)$, and variance $\sigma^2 = \mathbb{E}_F[(X-\mu)^2]$.

The skewness of a distribution is defined to be

$$\gamma(F) = \frac{\mathbb{E}_F\left[(X-\mu)^3\right]}{\sigma^3}.$$

(i) Define the empirical distribution function (ECDF) based on the sample $\{X_1, \ldots, X_n\}$, denoted $\widehat{F}_n(x)$. ***(3 marks)***

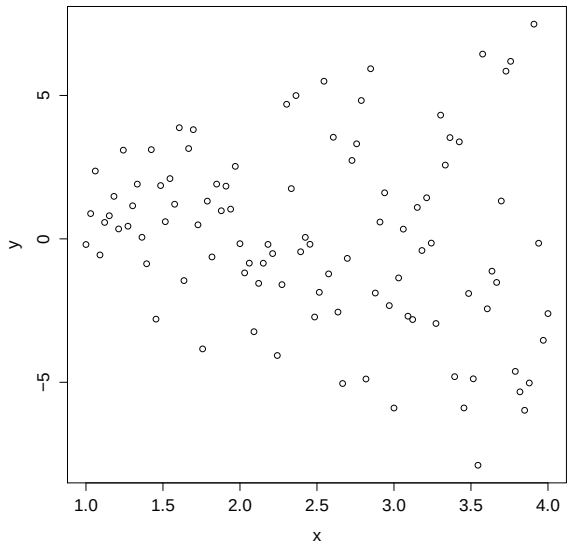
(ii) $\widehat{F}_n(x)$ is a random variable as it depends upon the random sample $\{X_1, \ldots, X_n\}$. What is the distribution of $\widehat{F}_n(x)$? ***(3 marks)***

(iii) Using the 'plug-in principle', find an estimator of the skewness of F based on the ECDF, i.e., calculate $\gamma(\widehat{F}_n)$. ***(5 marks)***

(iv) How would you produce a sample of size n from $\widehat{F}_n(x)$? ***(2 marks)***

(v) Describe a bootstrap procedure for estimating the standard error of $\gamma(\widehat{F}_n)$ and state how you would calculate a 95% confidence interval for $\gamma(F)$. ***(7 marks)***

2 A statistician is given observations (x_i, y_i) for $i = 1, \ldots, 100$. The data are plotted below.



Perhaps unwisely, they begin by fitting the following linear model:

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, 100$. The R output from their analysis is as follows

```
> fit <- lm(y~x)
> coef(fit)
(Intercept)           x
  2.3269881  -0.8744555
```

where `y` is the vector $(y_1, \ldots, y_{100})$ and `x` is the vector $(x_1, \ldots, x_{100})$

(i) The statistician then runs the following R commands:

```
> betas <- replicate(10^3,
+           {
+             x = sample(x, size=100, replace=F)
+             fit <- lm(y~x)
+             coef(fit)[2]
+           })
>
> beta_obs = coef(lm(y~x))[2]
> sum(abs(betas)>abs(beta_obs))
[1] 31
```

(a) State the null hypothesis being tested. ***(1 mark)***

(b) Explain the procedure that has been used to conduct the test. ***(2 marks)***

2 (continued)

(c) Give the result of the test. ***(2 marks)***

(d) Comment on the merits of this procedure compared to the t-test (which assumes that $\varepsilon_i \sim N(0, \sigma^2)$). ***(2 marks)***

(e) If there were only 5 observations rather than 100, what is the smallest possible p-value that could be obtained using this procedure? ***(2 marks)***

(ii) After discussion with their client, the statistician decides that a more appropriate model is

$$y_i = \beta x_i + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, x_i^2)$, for $i = 1, \ldots, 100$.

Explain how to perform a Monte Carlo test of size 0.01 of the hypothesis $H_0 : \beta = 0$ against the two-sided alternative $H_1 : \beta \neq 0$ using the weighted least squares estimator of β

$$\hat{\beta}(\mathbf{x}, \mathbf{y}) = \arg\min_{\beta} \sum_{i=1}^{100} \left(\frac{y_i - \beta x_i}{x_i} \right)^2$$

as a test statistic.

(4 marks)

(iii) Suppose f is a density function of two parameters, α and β, and that we observe iid data $\mathbf{x} = \{x_1, \ldots, x_n\}$ from $f(x; \alpha, \beta)$.

(a) Define the profile likelihood function for α. ***(2 marks)***

(b) Describe how you would find a 95% confidence interval for α based on the profile deviance function (assuming n is large). ***(5 marks)***

3 Consider the probability density function (pdf)

$$f(x) = \begin{cases} 3x^2 \mathrm{e}^{-x^3} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(i) Briefly describe how $U[0,1]$ random variables are generated on a computer. ***(2 marks)***

(ii) Find a function $g(U)$ such that if $U \sim U[0,1]$, then $X = g(U)$ is a random variable with pdf $f(\cdot)$. ***(4 marks)***

(iii) If we want to estimate $\mathbb{E}X$ when $X \sim f$, and if $\mathbb{V}\mathrm{ar}(g(U)) = 0.1054$, what is the minimum number of samples from f needed to estimate a 95% confidence interval for $\mathbb{E}X$ which has a width of less than 10^{-3}? Carefully justify your answer. ***(6 marks)***

(iv) Given that

$$\mathbb{C}\mathrm{ov}(g(U), g(1-U)) = -0.1050,$$

state an antithetic variables estimator we could use to estimate $\mathbb{E}X$, and calculate how many samples (evaluations of g) we now require to estimate the 95% confidence interval for $\mathbb{E}X$ (still of width 10^{-3}). ***(8 marks)***

4 (i) We wish to estimate

$$S = \mathbb{E}X^2 = \int x^2 f(x) \mathrm{d}x$$

using importance sampling.

(a) Explain how importance sampling, using an importance density $g(x)$, can be used to estimate S. ***(2 marks)***

(b) Suppose that

$$f(x) = 4x^3 \exp\left(-x^4\right) \qquad x \geq 0$$

and that we want to use a normal distribution as the importance density g. By considering a Taylor series expansion of $h(x) = \log f(x)$ about the mode of f, obtain the optimal mean and variance for the importance density.

Hint: find $m = \arg\max_x h(x)$, and expand $h(x)$ around m.

(6 marks)

(ii) Consider a random variable with the probability density function

$$f(x) = \frac{c}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathbb{I}_{x>a}$$

where $a > 0$ is a constant threshold, and c is a constant to be determined. Here, $\mathbb{I}_A$ is an indicator function which takes value 1 if A occurs and 0 otherwise. This is known as a truncated normal distribution, and can be seen as the distribution of a $N(0, \sigma^2)$ random variable conditional on the event $\{X > a\}$.

(a) A naïve way to sample from f is to generate $N(0, \sigma^2)$ random variables until a value is generated larger than a. Show that this approach would require, on average,

$$\frac{1}{\Phi\left(\frac{-a}{\sigma}\right)}$$

simulations from a $N(0, \sigma^2)$ distribution to generate each accepted value. Here, Φ is the CDF of a standard normal random variable

(3 marks)

Hint: recall that the mean of a Geometric(p) distribution is $1/p$

(b) Describe a rejection sampling algorithm to sample from f using a $N(\mu, \sigma^2)$ distribution as the proposal g, where μ is constrained to be positive, i.e. $\mu > 0$. Be sure to calculate the bounding constant $M = \sup_x \frac{f(x)}{g(x)}$. ***(6 marks)***

(c) What value of μ will make this algorithm most efficient?

(3 marks)

End of Question Paper