KENYATTA UNIVERSITY

UNIVERSITY EXAMINATIONS 2011/2012

SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF SCIENCE (STATISTICS AND COMPUTING)

SST 302: PROGRAMMING LANGUAGE FOR STATISTICS II

DATE: **Tuesday 3rd April 2012** **TIME:** **8.00a.m – 10.00a.m**

INSTRUCTIONS: Answer Question ONE [COMPULSORY] and any other TWO questions

1. (a) Define the following terms

   i. Random variable **(1mk)**

   ii. Monte Carlo methods **(1mk)**

   iii. Bootstrap methods **(1mk)**

   (b) Given a Cauchy random variate X with a cumulative distribution function

$$F(x) = \left\{ \frac{1}{2} + \frac{1}{\pi} arctan \left(\left(\frac{x-\mu}{\sigma} \right) \right) \quad 0 \leq \mu < \infty, \sigma^2 > 0 \right.$$

   Give an R code for generating random variables using the inverse transform method **(5mks)**

   (c) Consider a geometric probability mass function given by

$$P(X = j) = \begin{cases} pq^{j-1} & \text{for } j = 1, 2, \cdots \\ 0 & \text{Otherwise} \end{cases}$$

   i. Obtain the cumulative distribution function for Geometric probability mass function above **(3mks)**

   ii. Give an algorithm for simulating random variates from the distribution function obtained in (i) above **(4mks)**

   iii. Give an R code to implement the algorithm in (ii) above **(3mks)**

(d) Using the recursive property of the binomial mass function, illustrate an algorithm for generating random variates from a binomial mass function given by

$$P(X=j)=\begin{cases}\dfrac{n!}{(n-j)!j!}p^j(1-p)^{n-j} & \text{for } j=0,1,2,\cdots,n\\ 0 & \text{Otherwise}\end{cases}$$

. Also show the R code to implement the algorithm illustrated.

(10mks)

(e) Highlight any two advantages of bootstrap as a re-sampling technique to statistical inference

(2mks)

2. (a) Describe how to sample by rejection sampling from a density f using samples from a density g.

(3mks)

(b) On average, what proportion of observations from g will be accepted as observations from density f ? When will rejection sampling be efficient.

(2mks)

(c) Describe two advantages of rejection sampling over the inversion method for simulating from a distribution.

(2mks)

(d) Let X have a probability density function of the form

$$f(x)\propto\begin{cases}1-cos(x) & 0\leq x\leq\dfrac{\pi}{2}\\ 0 & \text{Otherwise}\end{cases}$$

i. Describe a rejection algorithm for simulating from the random variable X, based upon simulating over the rectangle $\left[0,\dfrac{\pi}{2}\right]\times[0,1]$. Calculate the acceptance probability of the algorithm

(6mk)

ii. Show how the method of inversion can be used to simulate from a random variable Y with

$$f(y)\propto\begin{cases}\left(\dfrac{y^2}{2}\right) & 0\leq y\leq\dfrac{\pi}{2}\\ 0 & Otherwise\end{cases}$$

Note that for all $0\leq x\leq\dfrac{\pi}{2}$, $1-cos(x)\leq\dfrac{x^2}{2}$. Use this fact to describe a rejection algorithm to simulate from X using Y. Calculate the acceptance probability of the rejection algorithm.

(7mks)

3. (a) For X, $E[X] = (8 - 4\pi + \pi^2)/(4(\pi - 2))$ and $E[X^2] = (48 - 6\pi^2 + \pi^3)/(12(\pi - 2))$. Given that $cov(exp(X), X)$, $cov(exp(X), X^2) > 0$, outline how control variates, using $E[X]$ and $E[X^2]$, can be used to estimate $\theta = E[exp(X)]$, based upon a sample $x_1, x_2, \cdots, x_n$ from X

(6mks)

(b) i. Describe the bootstrap as a way of estimating the statistical properties of a statistic $\varphi(X_1, X_2, \cdots, X_n)$ for an independent random sample $X_1, X_2, \cdots, X_n$. In particular, give expressions for the bootstrap estimate of the variance and bias

(8mks)

ii. For a random variable X, let $\theta = E[X]$. Suppose that $x_1, x_2, \cdots, x_n$ is a random (independent) sample from X. Then $\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is unbiased estimate of $E[X]$. Show that the variance of the bootstrap estimate of θ given the random sample $x_1, x_2, \cdots, x_n$ is $\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

(6mks)

4. (a) A time series x is assumed to come from a first order auto-regressive model

$$X_i = \rho X_{i-1} + \epsilon_i$$

where $2 \leq i \leq n$ for an unknown constant ρ. However the distribution of the errors $\{\epsilon_i\}$ is completely unknown except that it is reasonable to assume they are independent and identically distributed with mean 0

i. Find an expression for $\hat{\rho}$ where $\hat{\rho}$ is the least squares estimate of ρ

(5mks)

ii. Explain how you would simulate a bootstrap sample from the distribution of $\hat{\rho}$

(5mks)

iii. How would you use this bootstrap sample to test the hypothesis that $\rho = 0.5$? Give an outline of the procedure to be used.

(5mks)

(b) Give explicit details of how to simulate by inversion of the cumulative distribution function (cdf) from a random variable Z with probability density function,

$$g(z) = \begin{cases} \left(\dfrac{1215}{z^6}\right) & \text{for } (z > 3) \\ 0 & \text{otherwise.} \end{cases}$$

(5mks)

5. (a) Suppose that $Y_1, \cdots, Y_n$ is a random sample from a population and that θ is a population parameter that we would like to estimate. We have two potential estimators, $\hat{\theta}_1 = u_1(Y_1, \cdots, Y_n)$ and $\hat{\theta}_2 = u_2(Y_1, \cdots, Y_n)$. Carefully describe a simulation experiment to compare the performance of $\hat{\theta}$ and $\hat{\theta}$ as estimators of θ. You may assume that, for a given θ, we can generate (pseudo-) random instances from the population distribution. Are there any graphical techniques that you think might be useful in comparing pieces of output from the simulation?

(10mks)

(b) Consider a simple regression model

$$Y_t = \beta x_t + \varepsilon_t, \varepsilon_t \sim N\left(0, \sigma^2\right) \tag{1}$$

where $t = 1, \cdots, n$ and the ε_t are independent.

i. Write an R function that takes three arguments, the slope parameter β, the error standard deviation σ, and a vector of values of the explanatory variable x. The default values for β and σ should both be 1. The function should return simulated values $y_1, \cdots, y_n$ from model (1) where n is the length of the explanatory variable vector.

(2mks)

ii. Write an R function that uses the function in part (i) above and returns a vector of r instances of the parameter estimate $\hat{\beta}$.

(3mks)

iii. Write an R function whose inputs include a vector of possible β values and a vector of possible σ values and, for every possible combination simulates instances of $\hat{\beta}$ and outputs the simulated approximation to the mean square error. It should be clear from the output which pair of β and σ values each MSE value corresponds to

(5mks)