(*Knowledge for Development*)

KIBABII UNIVERSITY

UNIVERSITY EXAMINATIONS

2016/2017 ACADEMIC YEAR

FIRST YEAR FIRST SEMESTER

MAIN EXAMINATION

FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

COURSE CODE: **STA 872**

COURSE TITLE: **STATISTICAL COMPUTING**

DATE: 15/09/17 **TIME:** 2 PM -5 PM

INSTRUCTIONS TO CANDIDATES

Answer Any THREE Questions

TIME: 3 Hours

This Paper Consists of **4 Printed Pages**. Please Turn Over.

QUESTION ONE (20 MARKS)

a. Non-negative integers counts $Z_1,........,Z_n$ were generated independently from a Poisson distribution with some unknown mean λ, which we wish to estimate. Data was collected by writing each Z_i on a separate slip of paper, all of which were put in a box. Unfortunately, only now, when it is time to estimate λ, has it been realize that it's not possible to tell the difference between the number 6 and the number 9, because we don't know which way up each slip of paper is supposed to go. Therefore, all the numbers have been entered to the computer as 6, even though some of them may have actually been 9. Call the data entered $X_1,........,X_n$. The X_i are independent, but, due to the confusion of 6 and 9, the probability distribution for X_i is not Poisson, but is instead:

$$P(X_i = k/\lambda) = \begin{cases} e^{-\lambda}\lambda^k/k! & if \quad k \neq 6 \quad and \quad k \neq 9 \\ e^{-\lambda}\lambda^6/6! + e^{-\lambda}\lambda^9/9! & if \quad k = 6 \\ 0 & if \quad k = 9 \end{cases}$$

Show how to derive the formulas for an EM algorithm to find the maximum likelihood estimate for λ from $X_1,........,X_n$, with the unobserved data being the true counts associated with the values recorded as 6. Write an R function to implement this EM algorithm. It should take as arguments as a vector of X values, and the number of iterations of EM to do. Use some reasonable initial value for λ to begin the EM algorithm. (5 Marks)

b. Suppose that we want to obtain points, (x, y), that are uniformly distributed over the diamond shape with vertices at (0,1), (-1,0), (0,-1) and (1,0). Write an R function to do this using Gibbs sampling. You should use (0, 0) as the initial point to start the Markov chain, and then sample alternately for x and y, each of which will be vectors 1000 long, containing all the points from the chain (except the initial point). (5 Marks)

c. Suppose that we model a single real- valued point, x, as coming from a normal distribution with the mean μ and variance one. Suppose also that we use a prior distribution for μ that has the following density function over the real:

$$f(\mu) = \left(\tfrac{1}{2}\right)\exp(-|\mu|)$$

i. Write an R function called *met* that sample from the posterior distribution for μ using the Metropolis algorithm, with the proposal distribution for μ being normal with mean equal to the current value and variance one. Your function should take as arguments an initial value for μ and the number of transitions to do, and return a vector of values for μ after each transition. (5 Marks)

ii. Write R commands to use the *met* function from part (a) to find the posterior expected value of μ^2 given the observation $x = 1.5$. Use a starting value of zero for the Markov chain, and assume that the first 100 iterations should be discarded as ``burn-in" (not necessarily close to having the desired distribution). You should then estimate the expected value of μ^2 using 1000 iterations after the burn-in iterations.

(5 Marks)

QUESTION TWO (20 MARKS)

a. Suppose we have n binary (0/1) observations, $y_1,......., y_n$, that we model as being i.i.d. samples from a Bernoulli distribution with the probability of 1 being p. Suppose that 1/10 of the observations are 1s and 9/10 are zeros, so the maximum likelihood estimate of p is $\hat{p} = 1/10$. Suppose we use floating-point numbers of the form $0.d_1d_2d_3 \times 10^E$, where d_1, d_2, d_3 are decimal digits and E is an integer exponent in the range -100 to +100. For what values of n can we compute the probability of $y_1,...., y_n$ using maximum likelihood estimate $\hat{p}$ without the result underflowing to zero?

b. Suppose we try to solve the equation $x^4 - 81 = 0$ using Newton-Raphson iteration.

   i. How will we find the next guess at the solution, x^{t+1}, from the current guess, $x^{(t)}$?

   ii. Suppose we start from an initial value of $x^{(0)} = 5$. Here are the values found in the first four iterations:

$$x^{(1)} = 3.912$$

$$x^{(2)} = 3.2722427442656$$

x

$$x^{(3)} = 3.03212968848923$$

$$x^{(4)} = 3.00050709092522$$

   Estimate what value of $x^{(5)}$ will be, without actually doing the iteration.

QUESTION THREE (20 MARKS)

a. Suppose we numerically evaluate the integral

$$\int_0^1 x^4 dx$$

   Using the mid-point rule. Using 100 points, the approximation we get is 0.199983333625. Using 1000 points, the approximation we get is 0.199999833333363. Estimate what approximation we will get if we use the midpoint rule with 2000 points.

b. You've probably heard the rule that about 68% of values from a normal distribution will be within one standard deviation of the mean. Write an R program to verify this using the Trapezoidal Rule for integration, making use only of basic R facilities and the *dnorm* function.

QUESTION FOUR (20 MARKS)

Suppose we have n i.i.d (independent, identically-distributed) data points $x_1, \ldots\ldots, x_n$ that are real values in the interval (-1, +1). We model these observations as having the distribution on (-1, +1) with the following density function:

$$f(x) = (1 + \theta x)/2$$

Where θ is an unknown model parameter in the interval (-1, +1). We wish to find the maximum likelihood estimate for θ.

a. Write down the likelihood function, $L(\theta)$, and the log likelihood function, $l(\theta)$, based on the observations $x_1, \ldots\ldots, x_n$.
b. Write the R function so that it will compute the first derivative of the log of the likelihood for theta based on the observations in the vector x.
c. Write the R function so that it will compute the first derivative of the log of the likelihood for theta based on the observations in the vector x.
d. Write the R function so that it returns the maximum likelihood estimate for θ given the data vector x. Find the MLE using Newton iteration for *niters* iterations.

QUESTION FIVE (20 MARKS)

Suppose we have n i.i.d data points $x_1, \ldots\ldots, x_n$ that are positive real numbers with each having the distribution with density function

$$f(x) = \frac{1}{\theta(1 + x/\theta)^2}$$

Where θ is an unknown positive model parameter.

Derive the formulas needed to use Newton-Raphson iteration to find the maximum likelihood estimate of θ, and write an R program for it.