Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES (SIMS)
MASTER OF SCIENCE IN STATISTICAL SCIENCES
END OF SEMESTER EXAMINATION
STA 8204: COMPUTATIONAL STATISTICS

DATE: April, 2021 **TIME**: 3 Hours

INSTRUCTIONS

1. This examination consists of **FIVE** questions.
2. Answer Question **ONE (COMPULSORY)** and any other **TWO** questions.
3. You may use a **SIMPLE CALCULATOR**. No **MOBILE PHONES** in the exams room.
4. This examination will be administered in the **COMPUTER LAB**. No **PERSONAL LAPTOPS** are permitted into the exams room. **R markdown** is required for submission.

Question One (30 Marks)

(i) Use the information provided to answer (a) and (b). (8 marks)

(a) If $X \sim Exp(1)$, use the knowledge of probability integral transform to show that

$$X = -log\, U \sim Exp(1) \text{ if } U \sim U_{[0,1]}.$$

(b) Compare the random generations between uniform transform and the exponential distribution using 10000 random numbers.

(ii) Use the file "gapminder.csv" from the course website and read this into R. You're then required to perform the following analysis

(a) Obtain data set for only the year 1982. (3 marks)

(b) Obtain data set for the American countries in 1997. (3 marks)

(c) Add a column to the data set loaded in R with the new column being gdp which is computed as gdp=pop*gdpPercap (3 marks)

(iii) Consider minimizing the function

$$f(x) = |x - 3.5| + (x - 2)^2$$

on the interval $[0, 5]$ using the golden section search method. (4 marks)

(iv) Suppose we have the following artificial gene expressing values 2, 3, 1, 2 of Group 1, 8, 7, 9, 8 of Group 2, and 11, 12, 13, 12 of Group 3. By the use of an appropriate model implemented in R, is there any significance differences in means of the gene expressions? (5 marks)

(v) Describe the use of bootstrap in statistical estimation. (4 marks)

Question Two (20 Marks)

(i) Produce a table presenting various outcomes for the top 10 countries with highest MMR in 2017. Use the MMR point estimates for 2017 to select the 10 countries with the highest MMR. Then produce a table with the following information for each of the 10 countries: (10 marks)

(a) country name, sorted alphabetically

(b) the MMR point estimate for 2017, as well as the upper and lower bounds.

(c) the annual rate of reduction (ARR) for 2000-2017

(d) the estimated number of maternal deaths for 2017 as well as the upper and lower bounds.

(e) the share of maternal deaths as a proportion of the global total for 2017 for each of the 10 countries.

Output your results in a csv format. Save the file as Adm-Qs-2-i.

(ii) Suppose we have a method for generating a random variable Y having density function g(x). We can use this as basis for generating a random variable X having density function f(x). Describe the accept-reject algorithm that can be used to achieve this. (3 marks)

(iii) Determine an acceptance-rejection algorithm for the generation of a random variable from a Beta distribution (6, 3) using the uniform distribution (from 100000 random numbers) as the instrumental or candidate density. Generate the necessary plots showing the acceptance and rejection region. (7 marks)

Question Three (20 Marks)

(i) Suppose $X_1, \cdots, X_n$ are iid from Cauchy $(\theta, 1)$, with the following density:

$$f(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{1 + (x_i - \theta)^2}.$$

(a) Obtain the MLE for θ. (3 marks)

(b) Simulate the procedure in (a) a number of times to demonstrate MLE. Draw the log-likelihood functions based on each simulated datasets. (7 marks)

(ii) Suppose we observe nonzero iid samples $X_1, \cdots, X_n$ from a poisson distribution, whose probability function is given by

$$P(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda}{x!} \frac{1}{1 - \exp(-\lambda)}$$

Find the MLE of the above function by Newton-Raphson iteration in R. (10 marks)

Question Four (20 Marks)

(i) By including ten candidate explanatory variables (Weight, EngineSize, Horsepower, RPM, Rev.per.mile, Fuel.tank.capacity, Length, Wheelbase, Width, Turn.circle) and MPG.highway as the dependent variable from the Cars93 data in the MASS package, use the Akaike information criteria (AIC) to select the most appropriate model. (5 marks)

(ii) Does provision of airbags affect the maximum price that people are willing to pay for a car? Use Cars93 data in the MASS package to answer this question. (6 marks)

(iii) The following are results for an experiment in integrated circuit manufacture in which arsenic is deposited on silicon wafers. The factors are deposition time (A) and arsenic ow rate (B). Although both variables are quantitative, measurements are only taken at a high (labeled 1) and low level (labeled 0) of each. The purpose is to find out whether the factors have an effect and, if so, to estimate the sign and magnitude of the effect. The response is the thickness of the deposited layer.

Treatment combination	A	B	Thickness
(1)	0	0	14.037, 14.165, 13.972, 13.907
a	1	0	14.821, 14.757, 14.843, 14.878
b	0	1	13.880, 13.860, 14.032, 13.914
ab	1	1	14.888, 14.921, 14.415, 14.932

The figures for thickness are stored in a single column row by row in the file arsenic.dat. Load these data in R and answer this question. (9 marks)

Question Five (15 Marks)

(i) Consider dataset from the R package ISLR, which consists of the wages of 447 workers in 2005. use the bootstrap estimator to estimate the variance of the sample median. (8 marks)

(ii) Use the normal approximation to obtain a 95% confidence interval for the population median m in the mid-Atlantic states of the USA in 2005. (5 marks)

(iii) Load the Prestige data from library cars. Using the loess function, regress prestige on both the income and education levels of the occupations and examine the fitted regression surface graphically. (3 marks)

(iv) Address the statistical significance of each predictor in (iii) above by dropping it from the model and performing an approximate incremental F -test for the change in the residual sum of squares. (4 marks)

END