DS4043 - Final Examination part 2

Name: _____ Grade: _____

Coding Part

Before we start the exam, please install the following packages:

```
# install.packages("corrplot")
# install.packages("stats4")
# install.packages("MASS")
```

3. (20 points) In this question, we will use the dataset called *swiss*. This dataset is in our R package. We can simply call the data

```
data(swiss)
```

(a) (2 points) Show the number of observations and the names of variables in this data frame.

```
# Put your code here
```

(b) (4 points) Remove all observations from the data frame that the infant.mortality is less than 18.4 and call the new data frame *swiss.new*. Then use the *apply* function on this new data frame to calculate the mean and variance of each variable.

```
# Put your code here
```

(c) (4 points) Considering the data frame *swiss*, select the data such that the Catholic is among the 10% to 40% quantile range and make it a data frame called *M.catholic*. Print out the first 4 rows of *M.catholic*. (This question is Not Related to (b))

```
# Put your code here
```

(d) (4 points) Attach the data frame *swiss*. Plot the **probability density** histograms of *Fertility*, *Agriculture*, *Catholic* and *Infant.Mortality* where *Fertility* has 9 bins; *Agriculture*, *Catholic* and *Infant.Mortality* use the Sturges' method, Scott's method and Freedman-Diaconis' method respectively in the histograms function. (Note: you can directly use the arguments in the histogram function and no need to use formulas of these bandwidth methods) Display these four pictures as 2X2 by row in one plot.

```
# Put your code here
```

(e) (3 points) Attach the data frame *swiss*. Find the correlation matrix of *Fertility*, *Agriculture*, *Education* and *Infant.Mortality*. Then get the correlation plot and make the visualization method as **ellipse**, display the **full** matrix, the text label as **red** color, and show the diagonal elements.

```
library(corrplot)
# Put your code here
```

(f) (3 points) Attach the data frame *swiss*. Plot the Boxplot of *Fertility*, *Agriculture*, *Education* and *Infant.Mortality* in one boxplot picture.

```
# Put your code here
```

4. (10 points) Simulate a continuous Exponential-Gamma mixture. Suppose that the rate parameter λ has Gamma(r, β) distribution and Y has Exp(λ) distribution. That is, $(Y \mid \lambda) \sim f_Y(y \mid \lambda) = \lambda e^{-\lambda y}$; **in**

other words, the distribution of Y depends on λ and λ is generated from the Gamma(r, β) distribution.

(a) (6 points) Write a function to generate $n = 100$ random observations from this mixture with shape $r = 4$ and rate $\beta = 2$. (Hint: you should use gamma distribution to generate λ = rgamma(n,r,beta), and then use λ to generate Y). Print out the first 10 observations of the generated sample.

```
set.seed(10)
# Put your code here
```

(b) (4 points) Plot the density histogram of the generated sample using the exact Scott's method (**NOT** the one in the histogram function) and plot the sample density curve on the histogram.

```
# Put your code here
```

5. (14 points) We want to compute a Monte Carlo estimate of

$$\int_0^2 xe^{-x}\,dx$$

using the importance sampling method. We choose two importance functions $f_1 = 1/2$ which is the density of Uniform(0, 2) and $f_2 = \frac{e^{-x}}{(1-e^{-2})}, 0 < x < 2$. Show your estimation results and their corresponding standard errors using 10000 replications. **(Hint: Use the inverse transform method to generate sample from f_2)**

```
set.seed(39)
# Put your code here
m <- 10000
theta.hat <- se <- numeric(2)
```

6. (18 points) Consider the random variables $X_1, \ldots, X_n$ that are i.i.d. with a mixture density, i.e.

$$(1-p)N\left(\mu = 1, \sigma^2 = 4\right) + p\text{Exp}\left(3\right),$$

where Exp (3) is the exponential distribution with rate parameter 3. We have $\alpha = 0.05, p = 0.4$ and $n = 1000$. Let k_1 denote the sample kurtosis of random variable X defined as

$$k_1 = \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2\right)^2} - 3.$$

and we denote the excess kurtosis as Kt_1.

(a) (5 points) Generate a sample from this mixture normal density described above and compute k_1 of this sample.

```
set.seed(99)
# Put your code here
```

(b) (8 points) Compute the bootstrap and jackknife estimate of the bias and the standard error of the estimator k_1. Compare the results of bootstrap and jackknife. Note: we use $m = 1000$ for the bootstrap replicates.

```
# Put your code here
set.seed(95)
n <- 1000

## bootstrap method
m<- 1000
```

```
## jackknife method

## Compare Result
```

(c) (5 points) Consider the hypotheses for the excess kurtosis $H_0 : Kt_1 = 0$ vs $H_1 : Kt_1 \neq 0$. The test statistic is $G_2/se(G_2)$ where

$$G_2 = \frac{n-1}{(n-2)(n-3)}[(n+1)k_1 + 6]$$

and $se(G_2) = \sqrt{\frac{24}{n}}(1 - \frac{2}{n})$. The test statistic is approximately standard normal, so we reject the null hypothesis if the test statistic is beyond the critical z-value with significance level $\alpha = 0.05$ (note: two-side test). Use the Monte Carlo method to estimate **empirical power** of the hypotheses for the data generated as like in (a). The number of simulation is $m = 1000$.

```
# Put your code here
m <- 1000;  #num. repl.
n= 1000;
set.seed(75)
Gtests <- numeric(m)  #test decisions
```

7. (8 points) Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the probability density function

$$f(y \mid \theta) = \begin{cases} (\frac{2y}{\theta})e^{-y^2/\theta}, & y > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

(a) (3 points) Assume $\theta = 2$, use the inverse transform method to generate a sample with sample size $n = 200$ from this density function.

```
set.seed(27)
# Put your code here
```

(b) (2 points) Given the sample you generated in (a), use the analytical solution of MLE estimator of θ to estimate $\hat{\theta}$. **Note: you have already solved the expression of $\hat{\theta}$ in question 2, hand written part.**

```
# Put your code here
```

(c) (3 points) Given the sample you generated in (a), use the *mle* function in R to estimate $\hat{\theta}$. Choose the initial value of $\theta = 0.5$. (You can choose the lower bound = 0.1 and upper bound = 4)

```
# Put your code here
```