Statistical Computing 1, Stat 590 | Homework 4 | Name: ______________
Fall 2015 | | Prof. Erik B. Erhardt

Part I. (65 points) Do all calculations in LaTeX + R + knitr. Insert computer text output and graphics to support what you are saying. For this assignment, all R code should well commented and be visible (`echo=TRUE`) in the document where you have written it.

Goal: Construct a parametric bootstrap confidence interval for the coefficient of variation of waiting time to next eruption for the Old Faithful geyser using the rejection sampling method.

(15pts) **1.** The Old Faithful geyser dataset is in the `datasets` package.

```
library(datasets)
# ?faithful
# Old Faithful Geyser Data
# Waiting time between eruptions and the duration of the eruption for the
# Old Faithful geyser in Yellowstone National Park, Wyoming, USA.
#
# A data frame with 272 observations on 2 variables.
# [,1]   eruptions   numeric   Eruption time in mins
# [,2]   waiting     numeric   Waiting time to next eruption (in mins)
```

```
str(faithful)
```

```
## 'data.frame': 272 obs. of  2 variables:
##  $ eruptions: num  3.6 1.8 3.33 2.28 4.53 ...
##  $ waiting  : num  79 54 74 62 85 55 88 85 51 85 ...
```

```
head(faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```
summary(faithful)
```

```
##    eruptions        waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

(a) (5 pts) Plot the waiting time data and describe the pattern you see.

(b) (10 pts) A mixture distribution is of the form

$$f(x) = \sum_{i=1}^{k} \lambda_i f_i(x),$$

where λ_i is a proportional contribution of pdf $f_i(x)$ to the mixture $f(x)$.
Look at the help for the `normalmixEM()` function in the `mixtools` package.

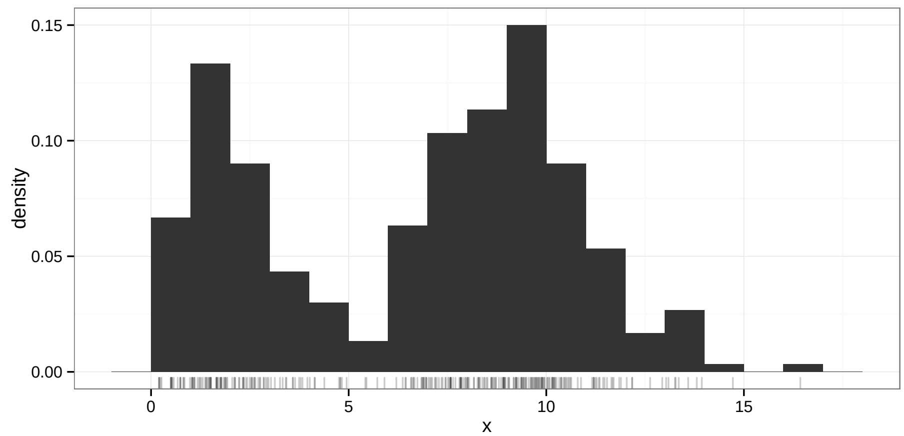
```
library(mixtools)
# ?normalmixEM

# generate some fake data to test the function
df.a <- data.frame(x = rnorm(100, mean = 2, sd = 1), dist = "A")
df.b <- data.frame(x = rnorm(200, mean = 9, sd = 2), dist = "B")
df.mix <- rbind(df.a, df.b)
df.mix$dist <- factor(df.mix$dist)

# inspect
summary(df.mix)
```

```
##        x           dist
##  Min.   : 0.2001   A:100
##  1st Qu.: 2.5599   B:200
##  Median : 7.5852
##  Mean   : 6.6410
##  3rd Qu.: 9.6965
##  Max.   :16.4273
```

```
# Inspect each distribution
library(plyr)
df.summary <- ddply(df.mix, 'dist', function(.subdf) {
    ## pull out the column of observations
    x <- .subdf$x
    data.frame( mean=mean(x), sd=sd(x), N=length(x))
})
df.summary
```

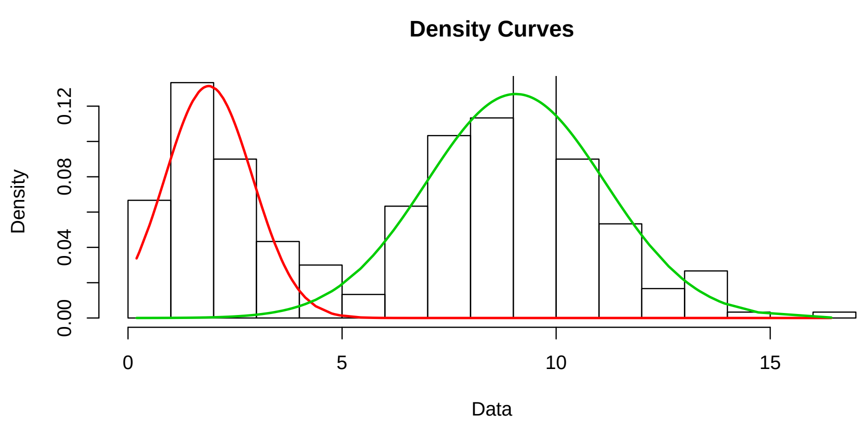
```
##   dist     mean        sd   N
## 1    A 1.852590 0.9875047 100
## 2    B 9.035222 2.1028740 200
```

```
# plot histogram of all data
library(ggplot2)
p <- ggplot(df.mix, aes(x = x))
p <- p + geom_histogram(aes(y=..density..), binwidth=1)
p <- p + geom_rug(alpha = 1/5)
p <- p + theme_bw()
print(p)
```



The parameters of the mixture distribution ($\lambda_i, \mu_i, \sigma_i$ for $i = 1, 2$) are well estimated with two normals when each component is based on a large sample size and the components are rather separate.

```
# estimate the mixture model parameters using the EM-algorithm
x.mix <- normalmixEM(df.mix$x)

## number of iterations= 67

x.mix[c("lambda", "mu", "sigma")]

## $lambda
## [1] 0.3369223 0.6630777
##
## $mu
## [1] 1.886775 9.056730
##
## $sigma
## [1] 1.022999 2.084844

# ?plot.mixEM  # plotting options
plot(x.mix, which = 2, breaks = 20) # plot density components
```



Use this strategy to estimate the mixture distribution parameters for the waiting time. Interpret the parameters.

(30^{pts}) **2. Simulating random deviates from a mixture distribution**

We will consider two strategies for drawing random deviates from the mixture distribution in the previous problem.

(a) (10 pts) Using the estimated parameters of the mixture distribution ($\hat{\lambda}_i, \hat{\mu}_i, \hat{\sigma}_i$ for $i = 1, 2$), one stategy is to draw a random deviate from each of the component distributions (Normal$(x|\hat{\mu}_1, \hat{\sigma}_1^2)$ or Normal$(x|\hat{\mu}_2, \hat{\sigma}_2^2)$) with probability proportional to their proportional contribution to the mixture ($\hat{\lambda}_1$ and $\hat{\lambda}_2$). Write code to simulate from the fitted mixture distribution using this strategy and plot a histogram based on a sample size equal to the original sample.

(b) (10 pts) The rejection sampling method can be used.
Set up the distributions needed for rejection sampling.

1. Using the estimated parameters of the mixture distribution ($\hat{\lambda}_i, \hat{\mu}_i, \hat{\sigma}_i$ for $i = 1, 2$) write a `function()` for the fitted density function $f(x)$ of the mixture distribution.
2. Determine a density which is easy to sample from, $h(x)$, and scale factor α to construct an envelope function $e(x) \equiv h(x)/\alpha$.
3. Show that this envelope function $e(x)$ is strictly not less than $f(x)$ over a sensible domain.

(c) (10 pts) Rejection sampling method, continued...
Perform the rejection sampling.

1. Draw x from the proposal distribution, $h(x)$.
2. Draw u from a uniform distribution.
3. Determine, using the rejection rule, whether to reject or accept x.
4. Repeat this until you have $n = 272$ accepted samples.
5. Plot a histogram of the samples.

(20^{pts}) **3. Parametric bootstrap**

(a) (10 pts) Using your results from #2, write a function to sample random deviates using the rejection sampling method. This function should have the following arguments: N, the number of samples; f.h, the proposal distribution function (and all associated parameters); alpha; and f.f, the target distribution evaluation function (and all associated parameters). It should return a vector of N deviates from the target distribution. A vectorized version of the above function is preferred. Half of the points will be based coding on style – write an efficient, organized, and well-documented function for full points.

(b) (10 pts) Use the above function to perform a parametric bootstrap to calculate $R = 10^4$ bootstrap values of the coefficient of variation and compute a central 95% CI.