STAT S 670 Exploratory Data Analysis - Homework #1

Ganesh Nagarajan
gnagaraj@indiana.edu

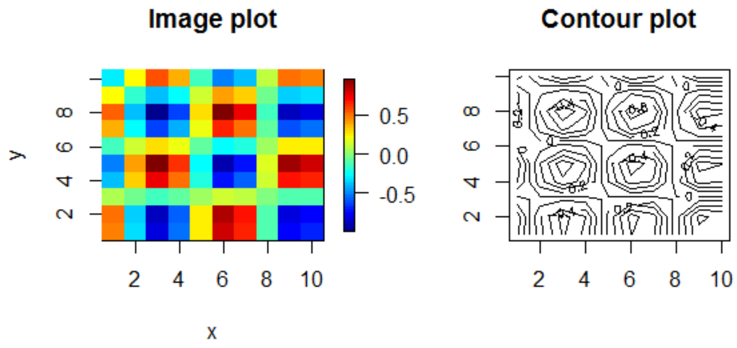
1 Solutions

1. Following is the R code to plot the given bivariate function $f(x, y) = cos(x)sin(y)$,

```
#Program 1
#Author: Ganesh Nagarajan
#Plots the image plot of function defined in bivariate.
library(fields)
biVariate <- function(x,y) {cos(x)*sin(y)}
plot3d <- function(xVector,yVector){
  x <- xVector
  y <- yVector
  if (length(x) == length(y)){
    plotMatrix <- outer(x,y,biVariate)
    #Replace all NaS in Plot Martix with Zeroes
    plotMatrix[is.na(plotMatrix)]<-0
    #Draw the Image Plot
    par(mfrow=c(1,2))
    image.plot(x,y,plotMatrix,main=”Image␣plot”)
    contour(x,y,plotMatrix,main=”Contour␣plot”)
  }
  else{
    print(”The␣size␣of␣the␣vectors␣don’t␣match␣each␣other”)
  }
}
#Plot the image map and plot the derivatives
plot3dDerivative<- function(xVector,yVector){
  plotMatrix <- outer(xVector,yVector,biVariate)
  image.plot(xVector,yVector,plotMatrix,main=”Image␣plot”)
  derivBiVariate <- deriv(~sin(x)*cos(y),  c(”x”,”y”),function(x,y){})
  gradient<-as.data.frame(attributes(
    derivBiVariate(xVector,yVector))$gradient)
  arrows(xVector,yVector,xVector+gradient$x,yVector+
           gradient$y,length = 0.25,col = ”red”)
}
```

For the sample input,

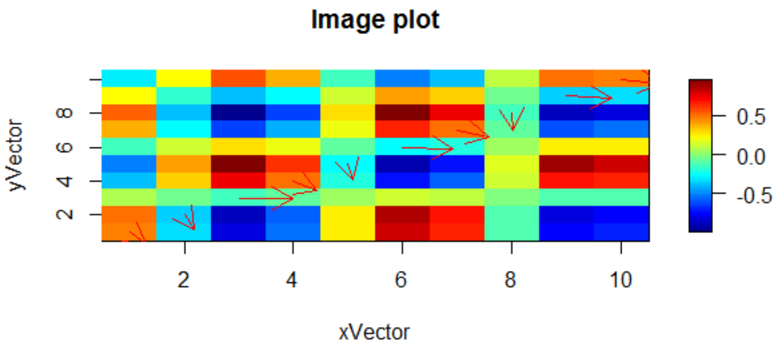
```
plot3d(1:10,1:10)
```



This function takes two vectors xVector and yVector as arguments which are in turn used to plot the perspective plot. It also checks if the length of xVector is equal to yVector to maintain consistency for the plot.

2. The arrows point to the gradient can be given as follows,

```
plot3dDerivative(1:10,1:10)
```



3. Stem and Leaf plots

   (a) i. For problem 2(c) Following is the manual calculation of stem leaf plots.

Depths	*Stem*	*Leaves*
2	3.	6 8
4	4*	1 2
6	4.	9 9
8	5*	0 1
10	5.	5 5
(2)	6*	3 3
9	7*	1 3
7	7.	6 9
5	8.	6 8
3	9.	0 9
1	10*	1

The above dataset need not be presented by the two stem per leaves model. The Dataset even if represented by normal scale it would not have been cluttered.

ii.
```
#Program 2
#Author: Ganesh Nagarajan
#Plots the stem and leaf plot for the given function.
stemleafplot<- function(listArray,scale){
  sortedArray <- sort(listArray)
  minArray <- min(listArray)
  maxArray <- max(listArray)
  count <- nrow(length)
  stem(listArray,scale)
}

#input vector from probem 2(b)
dataSet <- c(0.12,0.15,0.15,0.10,0.13,0.15,0.14,
             0.08,0.11,0.09,0.14,0.09,0.13,0.14,
             0.12,0.16,0.15,0.13,0.12,0.12,0.09)
#call the function with scale 0.25
stemleafplot(dataSet,0.25)
#call the function with scale 0.5
stemleafplot(dataSet,0.5)
#call the function with scale 1
stemleafplot(dataSet,1)
```

A. For scale 0.25, output is as follows,

```
The decimal point is 1 digit(s) to the left of the |
0 | 8999
1 | 01222233344455556
```

B. For scale 0.5, output is as follows,

```
The decimal point is 1 digit(s) to the left of the |

0 | 8999
1 | 012222333444
```

```
1 | 55556
```

C. For scale 1, output is as follows,
The decimal point is 2 digit(s) to the left of the |

```
08 | 0000
10 | 00
12 | 0000000
14 | 0000000
16 | 0
```

D. In this data set, scale 1 looses its accuracy, since would not be appropriate to represent the data.Scale of 0.25 is way too crowded for interpretation. By the rules, a two stem per leaf would be well appropriate for representing data and the scale of 0.5 would represent this in a concise manner.By the Rule, since n $<$ 100, $2\sqrt{21}$ rule can be used. Thus 9 stems are allowed and in all three scales, the stems have doesn't than 9 stems and hence is allowed allowed by the rule.

iii. This section describes the nature of distribution.

A. for 2(b) considering the scale of 0.5, the distribution seems to be near normal, mean is 0.1242857, median = 0.13. There are no outliers to the data set. The data set is almost symmetrical through the median. By the Rule, since n $<$ 100, $2\sqrt{n}$ rule can be used. Thus 9 stems are allowed and in all three scales. Since the stems are more than 9, the scale wouldn't be appropriate and not allowed by the rule.

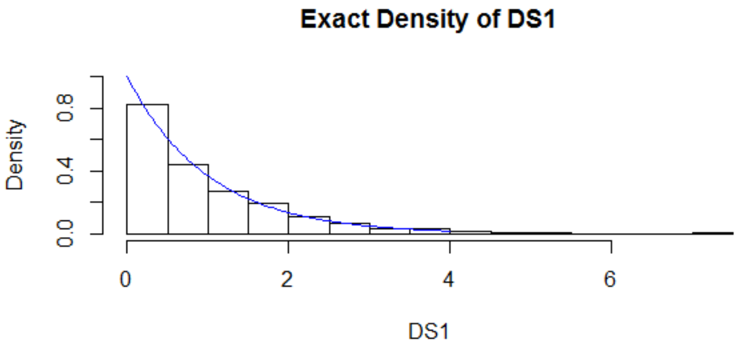
B. for 2(c), from the manually calculated stem leaf plot, there are no outliers and the data is equally distributed in all frequency. The data is almost symmetrical, however a skew is observed in the 10* portion of the stem leaf plot.

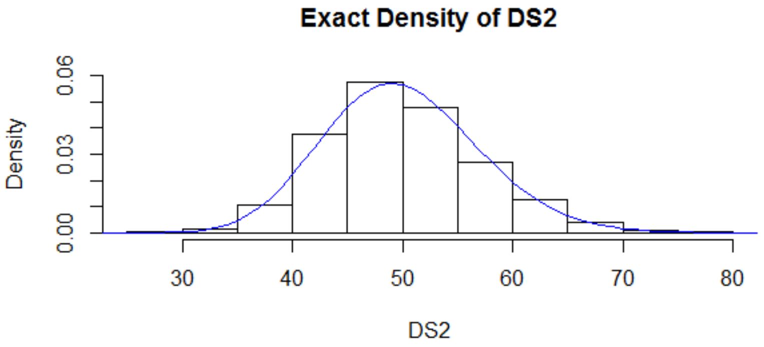
4. rgamma function and Kernel Density Estimates.
The R code for all plots in 4 question is as follows,

```
#Program 3
#Author: Ganesh Nagarajan
#Generate two Datasets using rGamma
set.seed(100)
dataSet1 <- rgamma(1000,shape = 1)
dataSet2 <- rgamma(1000,shape = 50)
#Draw the exact density function using curve.
hist(dataSet1,freq=F,ylim =  0+c(0,1),
      main=”Exact␣Density␣of␣DS1”,xlab = ”DS1”)
curve(dgamma(x,1),
        0,4,add=T,col=”blue”)
```

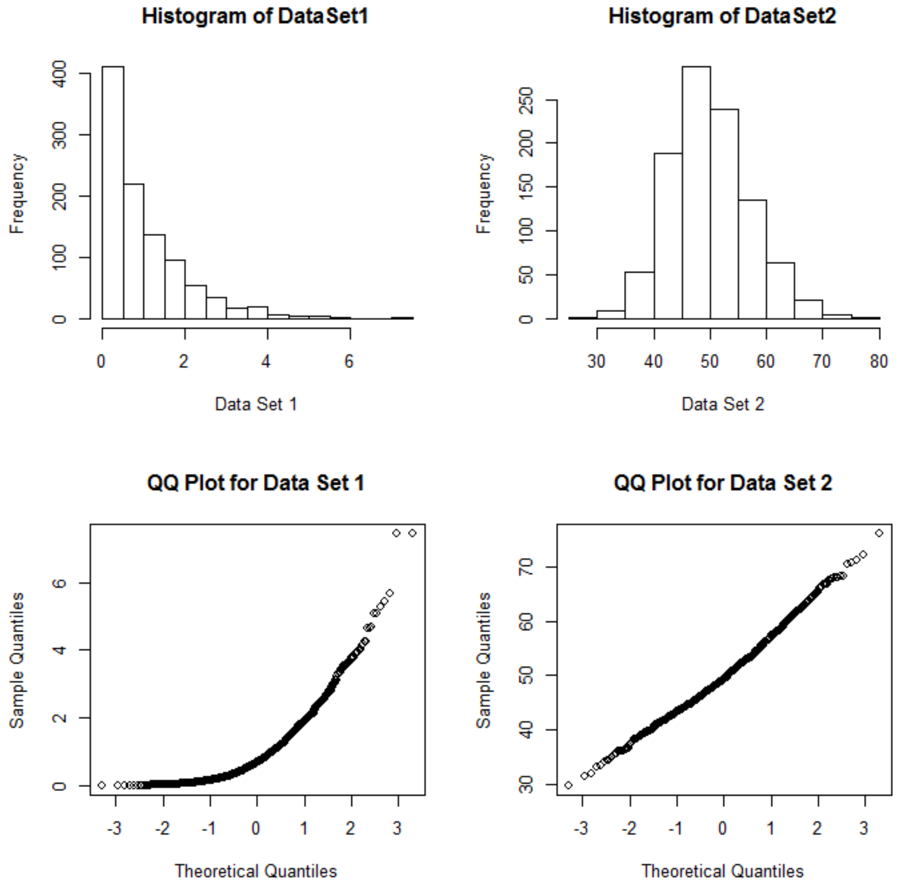
```
hist(dataSet2,freq=F,ylim =  0+c(0,0.06),
      main="Exact␣Density␣of␣DS2",xlab = "DS2")
curve(dgamma(x,50),
       0,100,add=T,col="blue")
#Configure the base plotting system for 2x2 plot
par(mfrow = c(2,2))
#Plot the Hitograms followed by the QQ plot
hist(dataSet1,main = "Histogram␣of␣DataSet1",xlab = "Data␣Set␣1")
hist(dataSet2,main = "Histogram␣of␣DataSet2",xlab = "Data␣Set␣2")
qqnorm(dataSet1,main = "QQ␣Plot␣for␣Data␣Set␣1")
qqnorm(dataSet2,main = "QQ␣Plot␣for␣Data␣Set␣2")
#Sumamry of the DataSet to determine the  mean-median relationship
summary(dataSet1)
summary(dataSet2)
par(mfcol = c(1,2))
#Plot the Histograms followed by the Kernel Distribution Plots for Data Set
hist(dataSet1,main = "Kernel␣Distribution␣Functions",
      xlab = "Data␣Set␣1",prob = 1)
lines(density(dataSet1,kernel = "gaussian"),col = 2,cex = 0.5)
lines(density(dataSet1,kernel = "rectangular"),col = 3,cex = 0.5)
lines(density(dataSet1,kernel = "triangular"),col = 4,cex = 0.5)
curve(dgamma(x,1),0,4,add=T,col=6)
legend(
  "topright",legend = c("Gaussian","Rectangular","Triangular","Exact␣Density
  col = c(2,3,4,6),lty = 1
)
#Plot the Histogram
hist(dataSet2,main = "Kernel␣Distribution␣Functions",
      xlab = "Data␣Set␣2",prob = 1)
lines(density(dataSet2,kernel = "gaussian"),col = 2)
lines(density(dataSet2,kernel = "rectangular"),col = 3)
lines(density(dataSet2,kernel = "triangular"),col = 4)
curve(dgamma(x,50),0,100,add=T,col=6)
legend(
  "topright",legend = c("Gaussian","Rectangular","Triangular","Exact␣Density
  col = c(2,3,4,6),lty = 1,cex = 0.5)
```

(a) Following is the exact density function for DataSet1 and DataSet2





(b) The QQ plot is as follows



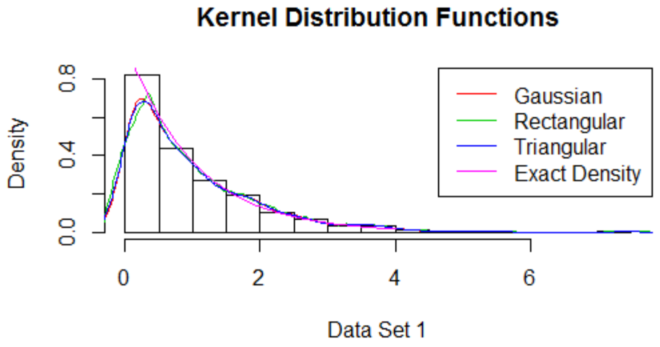
From the histogram and the qq plots, following can be inferred for dataset 1 and dataset 2.

i. For Dataset 1, the distribution is skewed right, as seen in the qq plot. The distribution does not seem normal and is evident from the proof that mean(0.998600) is far from median(0.692800). The distribution does not show any kind of symmetry.

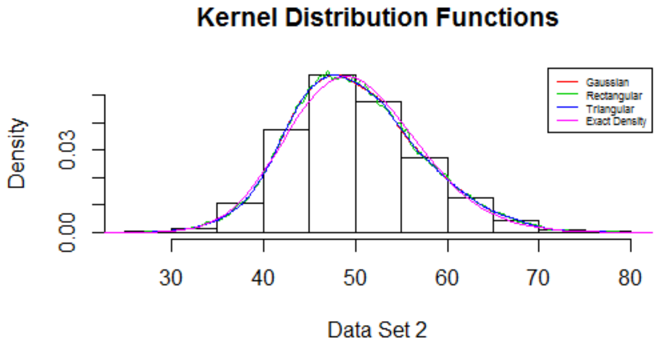
ii. For dataset 2, the distribution seems to be normal, this can be inferred from near straight line in qqplot. The distribution seems to suffer from fat tailed distribution to the right. Since this is a near-normal distribution mean(50) is near to median(49.44). There seems to be a symmetry through the median, however this is not perfect. This goes with the discussion in the class that the practical data sets and models seldom match with the perfect theoretical models and assumptions.

(c) Histograms are already available in the previous plot.

(d) For the purpose of this homework, only Guassian, Rectangular and Triangular Kernel Density measures are considered.
The Kernel Density distribution for dataset 1 is as follows,



The Kernel Density distribution for dataset 2 is as follows,



It can be inferred from the plots that for the data set 1, for rectangular method, there is a steep increase initially when compared to the Guassian and Triangular method. However for these data sets, Gaussian Distribution seems the most to represent the function.