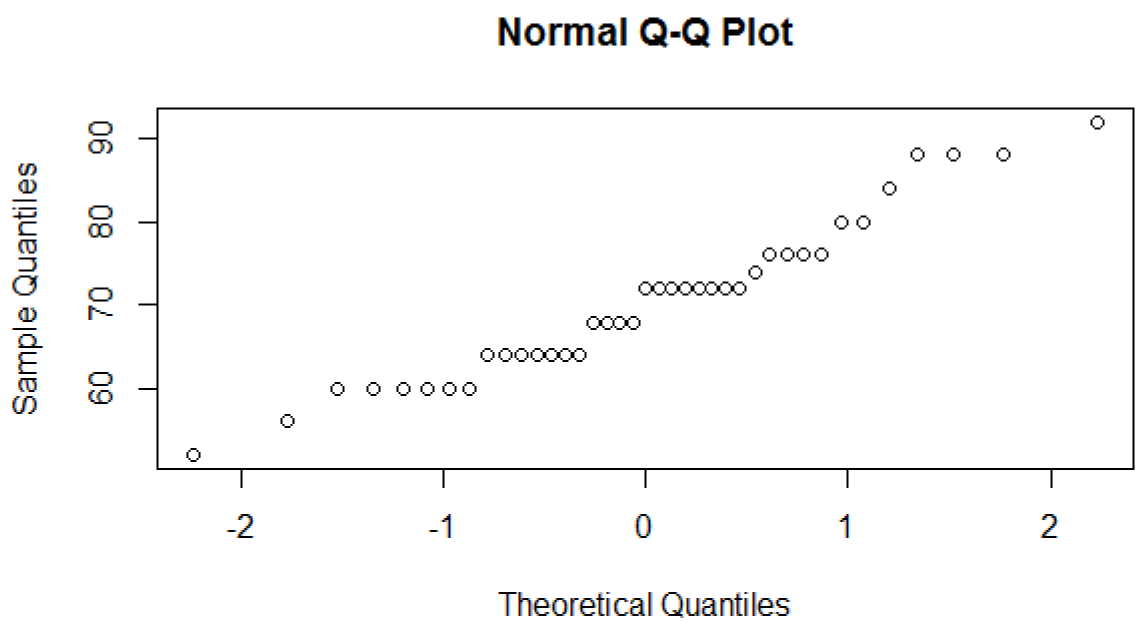
Problem Set 6

Problem 1

f) Construct a normal probability plot.

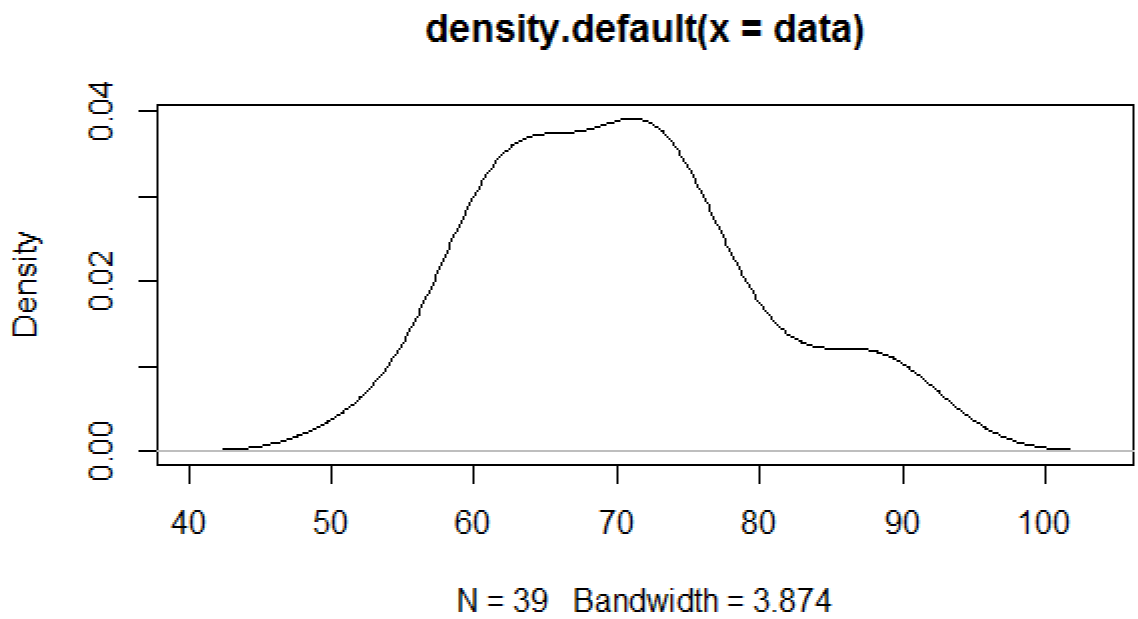
data = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/pulses.dat")

qqnorm(data)



g) Construct a kernel density estimate.

plot(density(data))



h) Do you think that this sample was drawn from a normal distribution?
Why or why not?

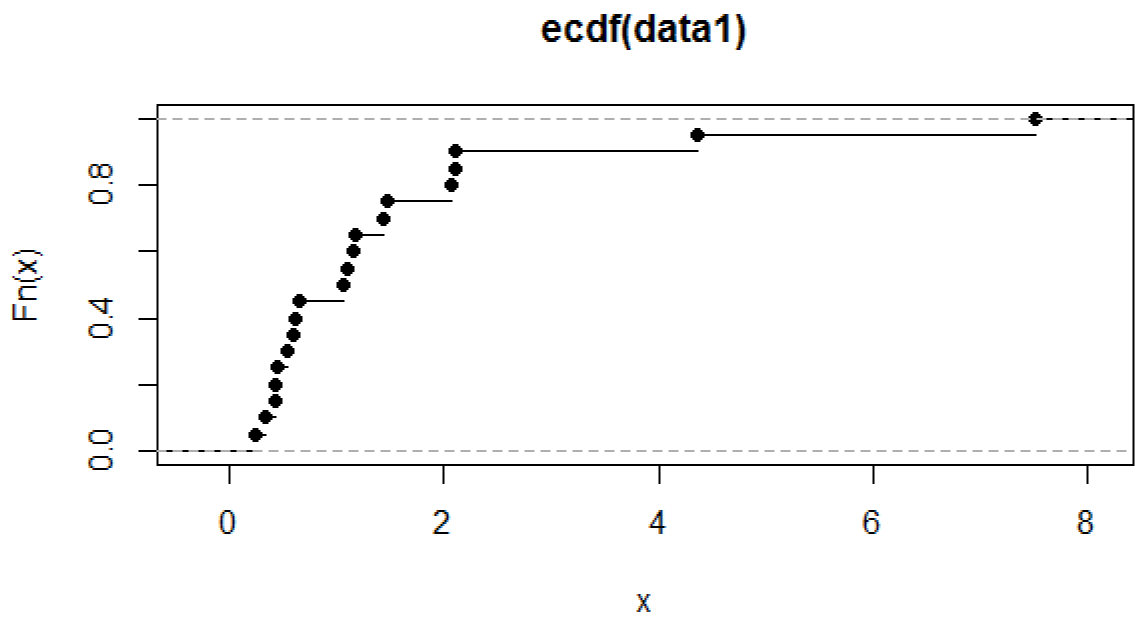
The data is scattered and does not form a straight line. Hence the data is not drawn from a normal distribution.

Problem 2

a) Graph the empirical cdf of _x.

```
data1=scan("Rdata.txt")

plot(ecdf(data1))
```



(b) Calculate the plug-in estimates of the mean, the variance, the median, and the interquartile range.

#mean

mean(data1) = 1.4876

#Variance

var=mean(data1^2)-(mean(data1))^2 = 2.787554

#Median

median(data1) = 1.076

#IQR

IQR(data1) = 1.10775

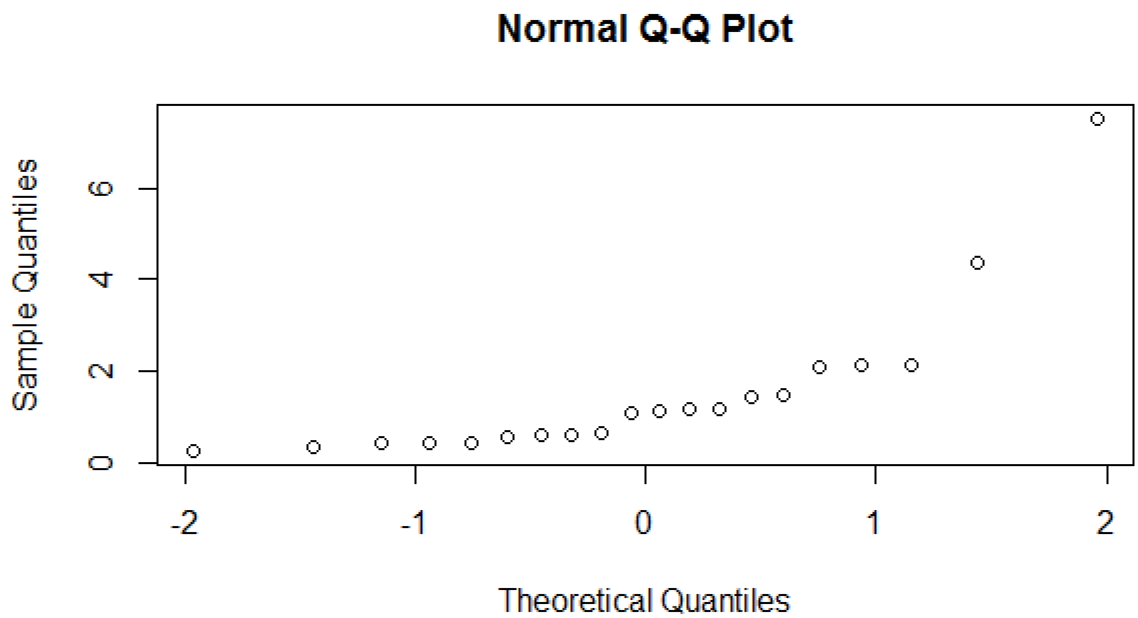
(c) Take the square root of the plug-in estimate of the variance and compare it to the plug-in estimate of the interquartile range. Do you think that _x was drawn from a normal distribution? Why or why not?

#Ratio of plugin estimate

IQR(data1)/sqrt(var) = 0.6634835

(d) Use the qqnorm function to create a normal probability plot. Do you think that _x was drawn from a normal distribution? Why or why not?

qqnorm(data1)



Its not a straight line. Hence data is not from Normal distribution.

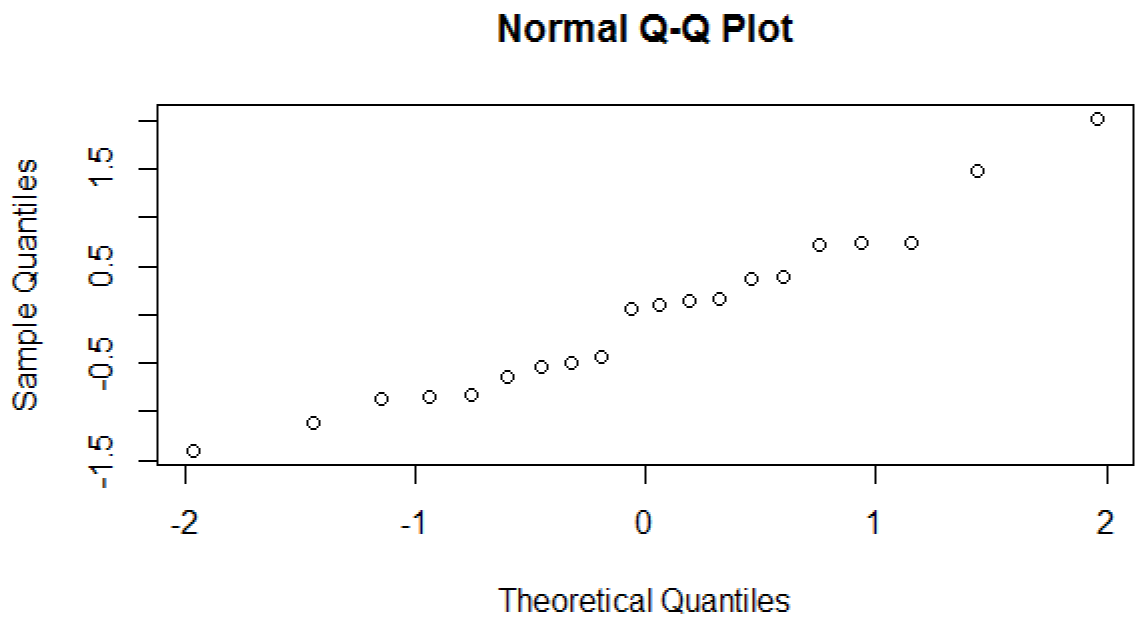
(e) Now consider the transformed sample _y produced by replacing each xi with its natural logarithm. If _x is stored in the vector x, then _y can be computed by the following R command:
> y <- log(x)
Do you think that _y was drawn from a normal distribution? Why or why not?

#qqnorm

#taking log of the data

y<-log(data1)

qqnorm(y)



The sample is very small. The ratio of interquantile range over the variance # gives us 1.35 which suggests that it might be from normal. However with this small data, we cannot conclude this.

####Question 3####

According to the text book, the law of averages suggests that as the size of the data increases, the average is almost the same. Since it's a fair coin, the probability will go towards 0.5. If there is some probability, 40 heads will give us some value. If the coin tosses increases, the probability will go towards 0.5.

"The Law of Averages formalizes our common experience that "things tend to average out in the long run." For example, we might be surprised if we tossed a fair coin n = 10 times and observed ¯X10 = 0.9; however,if we knew that the coin was indeed fair (p = 0.5), then we would remain confident that, as n increased, ¯Xn would eventually tend to 0.5. Clearly the sample 200 can be considered a substantial sample to expect that the same set of probability will be expected if we toss a coin."

Yes, the law of averages states as the size increases because we know it is a fair coin the probability will tend to get closer to 0.5, it does not mean expect 40 heads, but as the sample size increase the probability will get closer to .5

####Question 4####

(a) Approximate the probability that you will make a profit on your investment if you purchase a share of the value stock.

As its for market days , sample size n = 400 from the daily price fluctuations of the value stock.

Let $S_{400} = X_1 + X_2 + X_3 + ... + X_{399} + X_{400}$

So,

$E(S) = n * EX_i$

$= 400 * 0.01$

$= 4$

$Var(S) = n * Var X_i$

$= 400 * 0.01$

$= 4$

So, SD(S) = 4^0.5 = 2

Now, to see if we'll make profit on the investment, we will calculate:

$P(S > 0)$

We can write the following R code to calculate this:

1 - pnorm(0, 4 , 2) = 0.9772499

So,

$P(S> 0) = 0.977$

(b) Approximate the probability that you will make a profit on your investment if you purchase a share of the growth stock.

For Growth stock $EY_j = 0$ and $Var Y_j = 0.25$.

For sample size n = 400

$E(S_g) = n * EY_i$

$= 400 * 0$

= 0

Var(Sg) = n * Var Yi

= 400 * 0.25

= 100

So, SD(Sg) = 100^0.5 = 10

Now, to see if we'll make profit on the investment, we will calculate:

P(S> 0)

We can write the following R code to calculate this:

1 - pnorm(0, 0 , 10) = 0.5

Hence,

P(Sg > 0) = 0.5

(c) Approximate the probability that you will make a profit of at least $20 if you purchase a share of the value stock.

P(S>;= 20)

We can write the following R code to calculate this:

1 - pnorm(20, 4 , 2) = 6.661338e-16

(d) Approximate the probability that you will make a profit of at least $20 if you purchase a share of the growth stock.

Here we need to find out P(Sg>=20)

1 - pnorm(20, 0 , 10) = 0.02275013

(e) Assuming that the growth stock fluctuations and the value stock fluctuations are independent, approximate the probability that, after 400 days, the price of the growth stock will exceed the price of the value stock.

Let Sy = Sg – S

E[Sy] = E[Sg] – E[S]

= 0 – 4

=-4

Var(Sy) = Var(Sg) + (-Var(S))

= 4 + 100

=104

Sd(Sy) = sqrt(Var(Sy)

= 10.2

We need to find P(Sy>0)

Which can be calculated in R by writing:

1- pnorm(0, -4, 10.2) = 0.3474433

Hence P(Sy>0)= 0.347

####Question 5####

a) Find EX.

EX = Summation(x.fx)=(.3*-2+.6*-1+.1*12) = -.6-.6+1.2 = 0

b) For Var(X)

We have EX=0, EX^2 = .3*4+.6*1+.1*144 = 1.2+.6+14.4 = 16.2

Var(X)= EX^2 - (EX)^2 = 16.2-0=16.2

c) EX(X-bar) = 0 (Its the same as expected value)

d) Variance(X-bar) = var(x)/n = 16.2/n

e) for n=100, Variance(X-bar)= 16.2/100 = .1620

Hence standard deviation sd = sqrt(0.162) = 0.402

P(X-bar)>0.5 = 1-pnorm(0.5,mean=0,sd=0.402) = 0.1067

####Question 6####

```
hdata<-read.csv("data.csv",header = TRUE)

data2<- rep(hdata$Household, hdata$size.Number.of.households)

fun <- function(x) {

  c(mean = mean(x), sd = sd(x))

 }
```

- fun(data2)

```
mean       sd
2.500000 1.410638
```

a) Lacking any other information, our best estimate for the population mean household size is the sample mean. What is the sample mean of our data?

mean is 2.5

(b) What is our estimate for the standard deviation of household sizes?

sd = 1.41

(c) What is the estimated standard error of the sample mean? (That is, based on our answer to (b), what is our estimate for the standard deviation of the distribution of the sample mean?)

sample mean 1.41/sqrt(100) = .141

(d) Our error is the difference between the sample mean and the population mean. Using the normal distribution, and the approximate probability that the absolute value of the error in a survey of this form and size is less than 0.5.

pnorm(0.5, 0, 0.014) - pnorm(-.5, 0, 0.014) = 1

(e) Can we be reasonably sure that the average household size for all U.S. households is between 2 and 3?

Yes this is between 2 and 3 because the error is less than 0.5