problem set 6

Dilip Nikhil Francies

2023-10-13

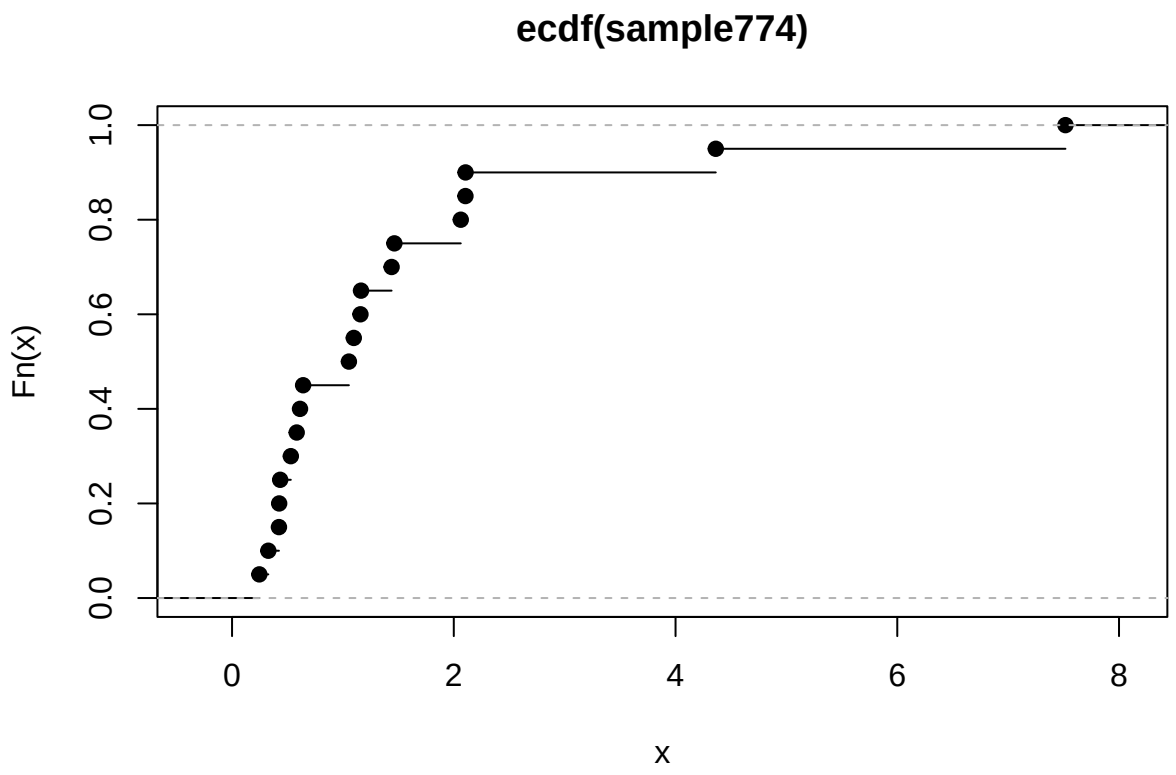
Question 2 :

```
#import data
sample774 <- scan("https://mtrosset.pages.iu.edu/StatInfeR/Data/sample774.dat")
sample774
```

2a : Draw the ecdf of xbar

```
##  [1] 0.246 0.327 0.423 0.425 0.434 0.530 0.583 0.613 0.641 1.054 1.098 1.158
## [13] 1.163 1.439 1.464 2.063 2.105 2.106 4.363 7.517
```

```
#plot ecdf
plot(ecdf(sample774))
```



```
plug_mean_x <- mean(sample774)
print(paste("Plug in mean of sample is ",plug_mean_x))
```

2b : Calculate the plug-in estimates of the mean, the variance, the median, and the IQR range

```
## [1] "Plug in mean of sample is  1.4876"
```

```
plug_var_x <- var(sample774)
print(paste("Plug in variance of sample is ",plug_var_x))
```

```
## [1] "Plug in variance of sample is  2.9342672"
```

```
plug_median_x <- median(sample774)
print(paste("Plug in median of sample is ",plug_median_x))
```

```
## [1] "Plug in median of sample is  1.076"
```

```
plug_iqr_x <- IQR(sample774)
print(paste("Plug in IQR of sample is ",plug_iqr_x))
```

```
## [1] "Plug in IQR of sample is  1.10775"
```

```
#square root of plug in variance
sqrt_plug_x <- sqrt(plug_var_x)
print(paste("The square roor of plug in variance",sqrt_plug_x ))
```

2c : Take the square root of the plug-in estimate of the variance and compare it to theplug-in estimate of the interquartile range. Do you think that x was drawn from a normal distribution? Why or why not?

```
## [1] "The square roor of plug in variance 1.71297028579015"
```

```
print(paste("The plug in iqr of x",plug_iqr_x))
```

```
## [1] "The plug in iqr of x 1.10775"
```

For a normal distribution, we know that the IQR is given by the formula IQR = 1.349 * SD.

```
norm_iqr = 1.349 * sqrt_plug_x
norm_iqr
```

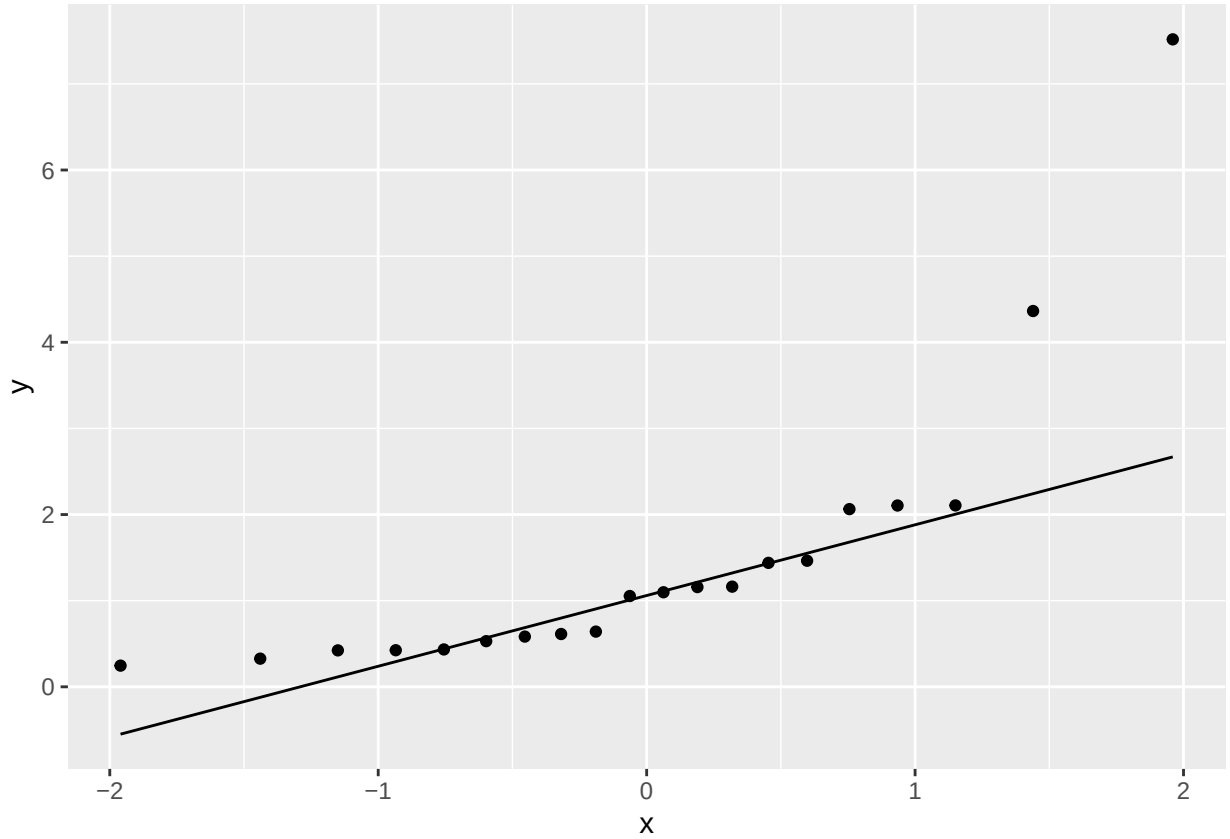
```
## [1] 2.310797
```

If the sample was of normal distribution, the plug in IQR should be close to the IQR that is calculated through the formula. Let's plot a qqplot to confirm.

```
#qqplot
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

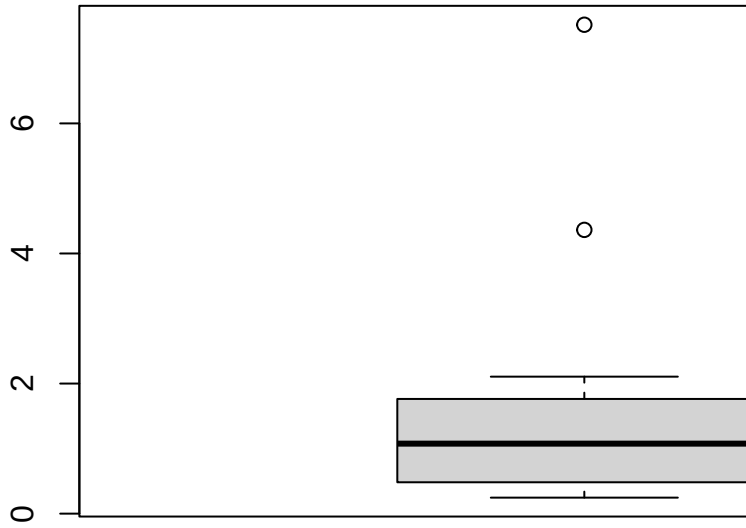
```
ggplot(data.frame(sample774),aes(sample=sample774)) + stat_qq() + stat_qq_line()
```



As one can see, the points do not align with the normal qq line, hence we can conclude that the sample x was not drawn from a normal distribution.

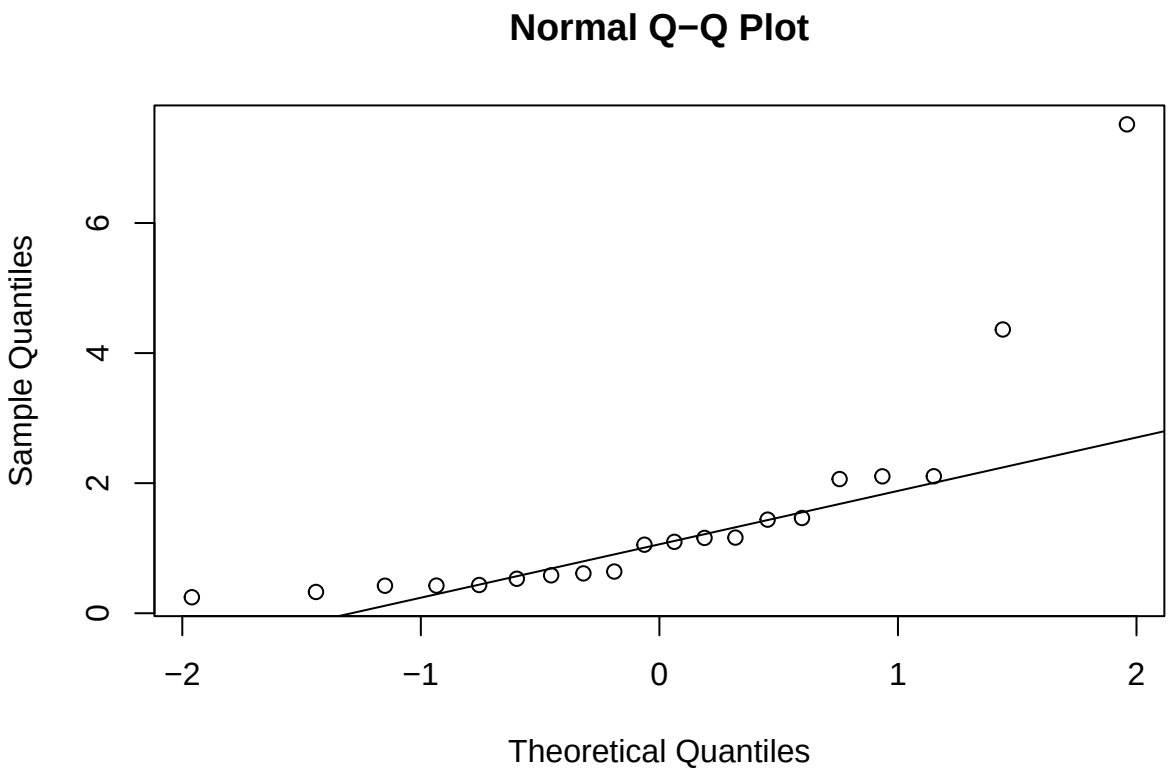
```
boxplot(sample774)
```

2d: 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?

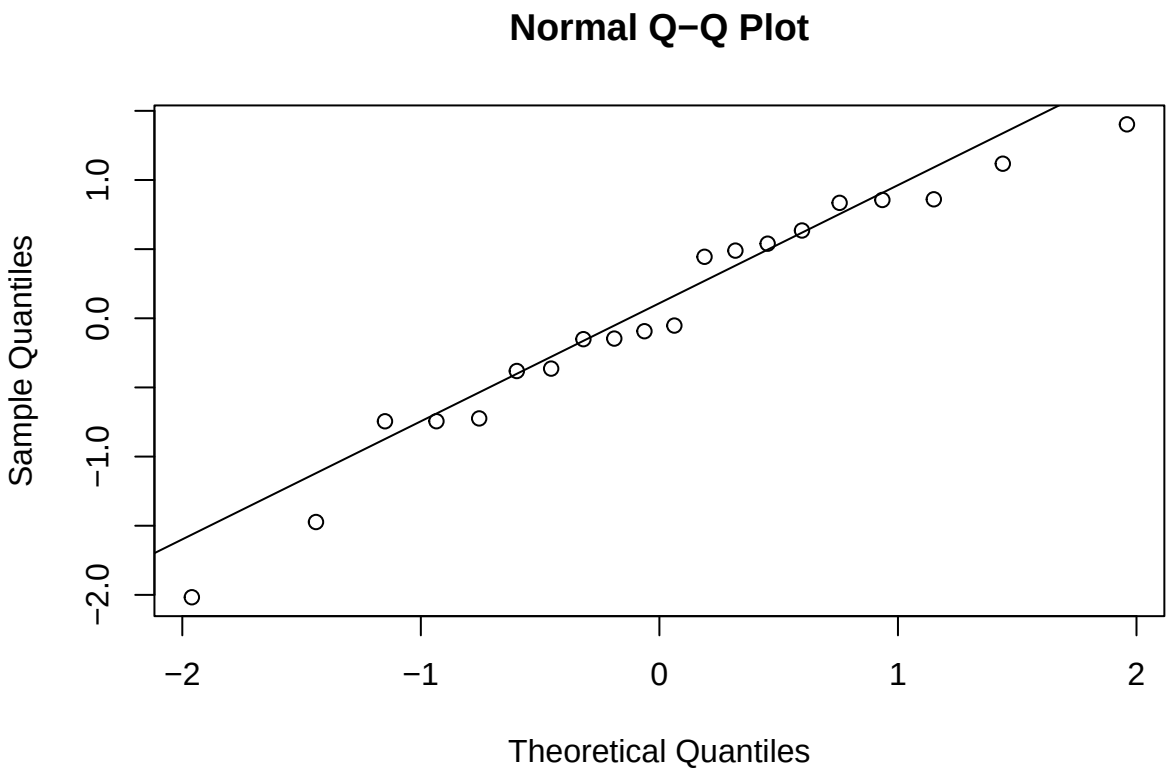
```
qqnorm(sample774)
qqline(sample774)
```



The graph clearly does not look like it resembles the empirical distribution of a normal distribution, because if it did, then majority of points would fall on the straight line. Even if there were minor sampling variation, the departure from the linearity should be minimal. Also from the box plot, we can see that there are plenty of outliers beyond the upper whisker. We can expect 7 outliers per 1000 observations in a normal distribution[chapter 7 trosstet] ,which is not the case here. hence, this sample is not likely to have been drawn from a normal distribution.

```
y <- -log(sample774)
qqnorm(y)
qqline(y)
```

2e : Now consider the transformed sample y produced by replacing each x i 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?



Though transformations such as logarithms and square root successfully convert non-normal distribution to normal, unfortunately it is not the case with this distribution. Though it does have linear trend, the points do not fall on the normal QQ line. Hence, "y" also does not look like it was drawn from normal distribution.

Q3. Consider an urn that contains 10 tickets, labeled { 1 , 1 , 1 , 1 , 2 , 5 , 5 , 10 , 10 , 10 } . From this urn, I propose to draw (with replacement) n = 40 tickets. I am interested in the sum, Y , of the 40 ticket values that I draw.

```
#define the values of tickets as a vector

tickets <- c(1,1,1,1,2,5,5,10,10,10)

#create a function

urn.model <- function(n) {            # n = sample size of your simulation
  sim <- sample(tickets,n,replace  = TRUE)    #replacement = True as mentioned
  sum_sim <- sum(sim)
  return (sum_sim)
}

#calculate the sum of 40 samples drawn from the urn through the function defined above

sum_sample <- urn.model(40)
print(paste("Sim 1: Sum of 40 tickets drawn from the urn with replacement is ",sum_sample))
```

3a: Write an R function named urn.model that simulates this experiment, i.e evaluating urn.model is like observing a value, y , of the random variable Y.

```
## [1] "Sim 1: Sum of 40 tickets drawn from the urn with replacement is  193"
```

```
#
```

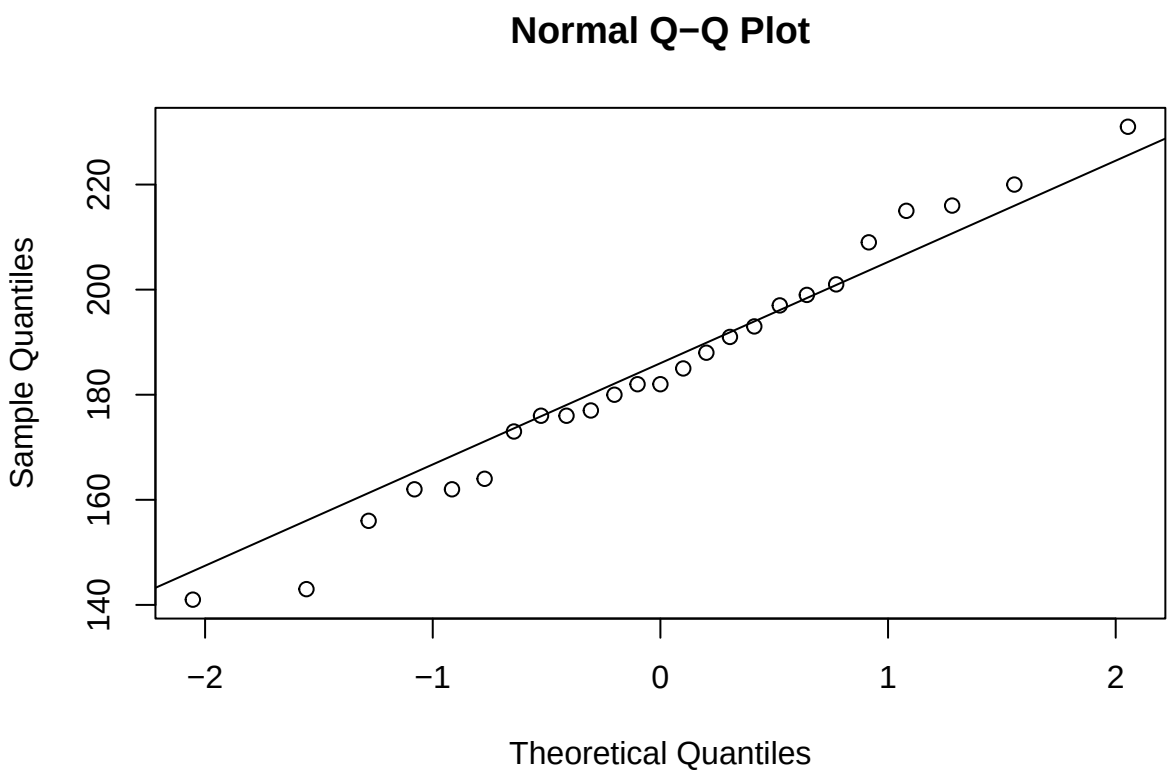
```
#loop through the urn.model 25 times to create the distribution
sample_size <- 25
sum_sample_sim <- numeric(sample_size)   #define an empty vector to append the values from simulation
for (i in 1:sample_size) {   # iterate 25 times
  sum_sample_sim[i] <- urn.model(40)
}

sum_sample_sim
```

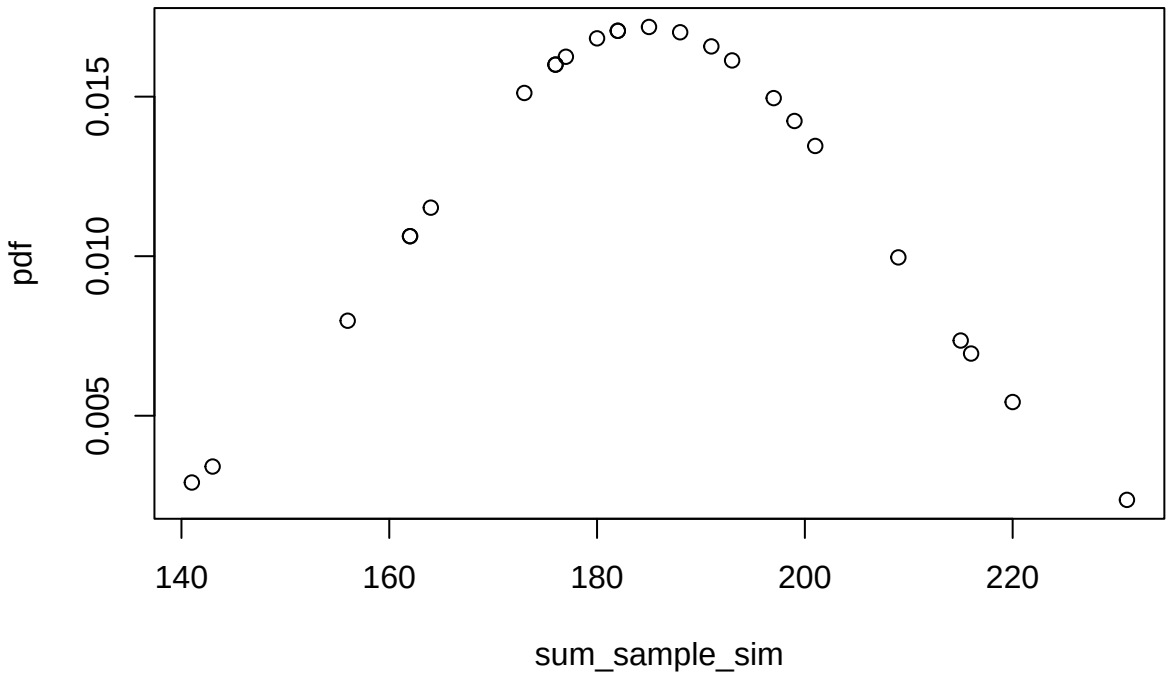
3 b) Use urn.model to generate a sample, y = { y 1 , . . . , y 25 } , of n = 25 observed sums.The random variable Y is discrete. Does it appear that the distribution of Y can be approximated by a normal distribution? Why or why not?

```
##  [1] 193 173 177 197 215 182 188 176 220 156 185 182 191 216 141 164 231 162 162
## [20] 199 201 209 176 180 143
```

```
#QQ plot to check for normality
qqnorm(sum_sample_sim)
qqline(sum_sample_sim)
```



```r
#calculate pdf
pdf<- dnorm(sum_sample_sim,mean(sum_sample_sim),sd(sum_sample_sim))
#plot the graph
plot(sum_sample_sim,pdf)
```



After looking at the qqplot and probability distribution plot, one can safely conclude that the distribution of sums of tickets does represent a normal distribution. Though the curve in the qqplot does deviates a tiny bit from linearity , it is expected because of small sample sizes.

```r
1-pnorm(105,100,2.236)
```

Question 5 code[Answer written]:

```
## [1] 0.01267143
```