MA677 Final Project

Keliang Xu

2022-05-11

Exercise

First, I try to solve the exercise problem in book. For exercise 4.25,

4.25

```
# pdf function
pdf <- function(x,a=0,b=1) dunif(x,a,b)
# cdf function
cdf <- function(x,a=0,b=1) punif(x,a,b,lower.tail = FALSE)

#  the distribution of the order statistics in Exercise 2.4
integrand <- function(x,r,n) {
  x * (1 - cdf(x))^(r-1) * cdf(x)^(n-r) * pdf(x)
}

# get expectation
E <- function(r,n) {
  (1/beta(r,n-r+1))*integrate(integrand,-Inf,Inf, r,n)$value
}

# approximation function
medianprrox<-function(k,n){
  return((k-1/3)/(n+1/3))
}
# for n=5
E(2.5,5)
```

```
## [1] 0.4166667
```

```
medianprrox(2.5,5)
```

```
## [1] 0.40625
```

```
# for n=10
E(5,10)
```

```
## [1] 0.4545455
```

```
medianprrox(5,10)
```

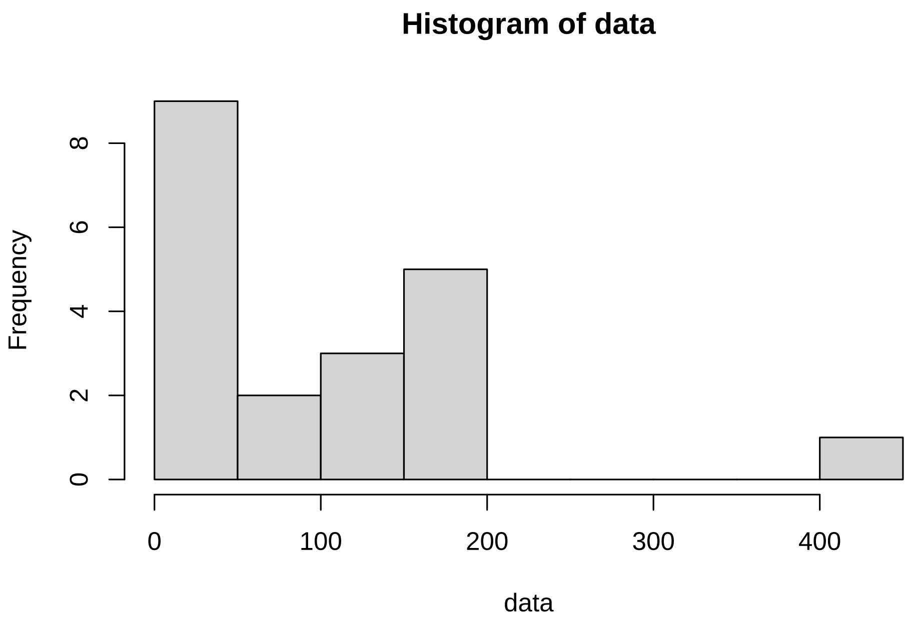
```
## [1] 0.4516129
```

As we can see, the results showed that they are quit similar.

4.39

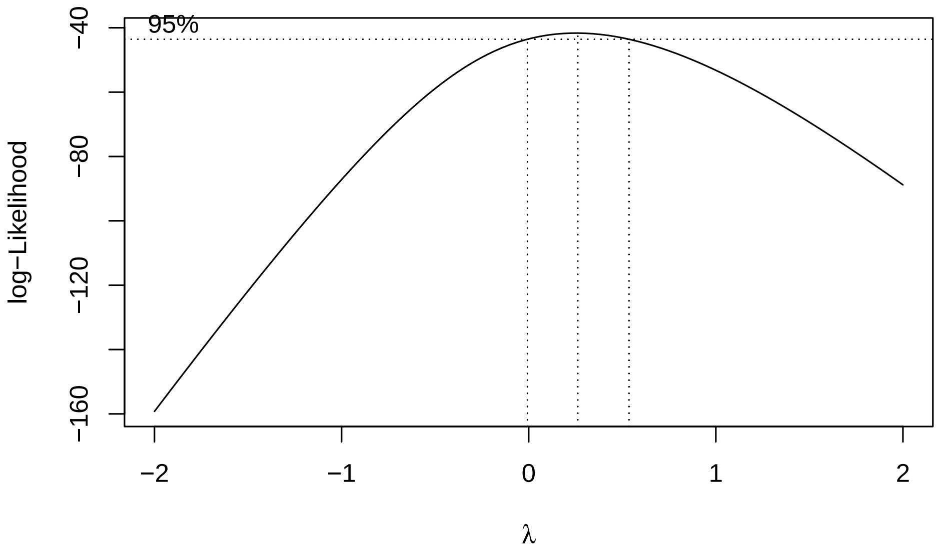
First, I load in the data of the average adult weight(in kg) of 28 species of animals.

```
data<-c(0.4,1.0,1.9,3.0,5.5,8.1,12.1,25.6,50.0,56.0,70.0,
        115.0,115.0,119.5,154.5,157.0,175.0,179.0,180.0,406.0)
hist(data)
```



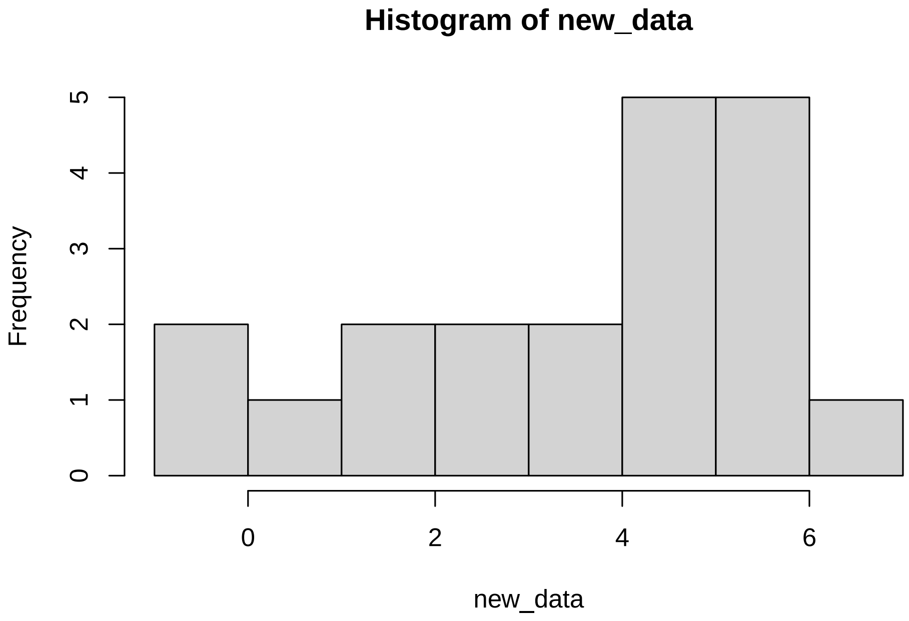
According to book, I try box-cox transformation which R function is boxcox(). And for value of new data, I get the answer from the network. https://nickcdryan.com/2017/04/19/the-box-cox-transformation/ And all the code I follow by this link of R code. https://r-coder.com/box-cox-transformation-r/

```
# install.packages(MASS)
library(MASS)
b <- boxcox(lm(data ~ 1))
```



As the previous plot shows that the 0 is inside the confidence interval of the optimal λ and as the estimation of the parameter is really close to 0 in this example, the best option is to apply the logarithmic transformation of the data.

```
# Transformed data
new_data <- log(data)
# Histogram
hist(new_data)
```



Now the data looks more like following a normal distribution, but we can also perform, for instance, a statistical test to check it, as the Shapiro-Wilk test:

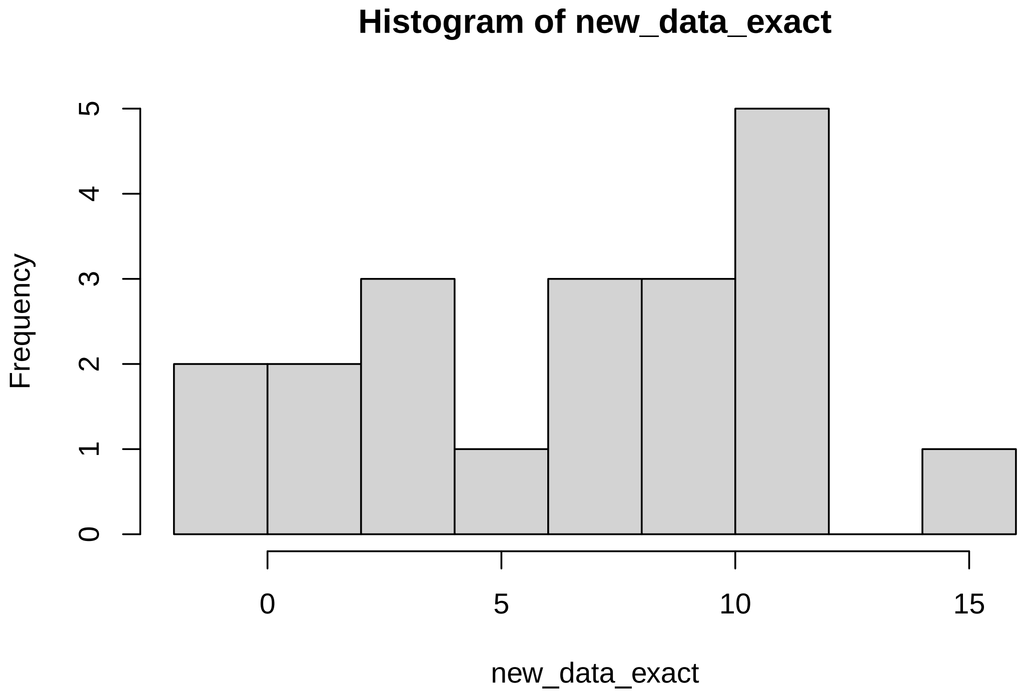
```
shapiro.test(new_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_data
## W = 0.89136, p-value = 0.02849
```

As the p-value is smaller than the usual levels of significance (1%, 5% and 10%) we need to reject the null hypothesis of normality.

So we need to extract the exact lambda using the following code.

```
## Extracting the exact lambda
la <- b$x[which.max(b$y)]
new_data_exact <-(data^la- 1)/la
hist(new_data_exact)
```



4.27

First, I load in the data in the book

```
Jan<-c(0.15,0.25,0.10,0.20,1.85,1.97,0.80,0.20,0.10,0.50,0.82,0.40,1.80,0.20,1.12,1.83,
       0.45,3.17,0.89,0.31,0.59,0.10,0.10,0.90,0.10,0.25,0.10,0.90)
Jul<-c(0.30,0.22,0.10,0.12,0.20,0.10,0.10,0.10,0.10,0.10,0.10,0.17,0.20,2.80,0.85,0.10,
       0.10,1.23,0.45,0.30,0.20,1.20,0.10,0.15,0.10,0.20,0.10,0.20,0.35,0.62,0.20,1.22,
       0.30,0.80,0.15,1.53,0.10,0.20,0.30,0.40,0.23,0.20,0.10,0.10,0.60,0.20,0.50,0.15,
      0.60,0.30,0.80,1.10,0.2,0.1,0.1,0.1,0.42,0.85,1.6,0.1,0.25,0.1,0.2,0.1)
```

a Compare the summary statistics for the two months.

```
summary(Jan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1875  0.4250  0.7196  0.9000  3.1700
```

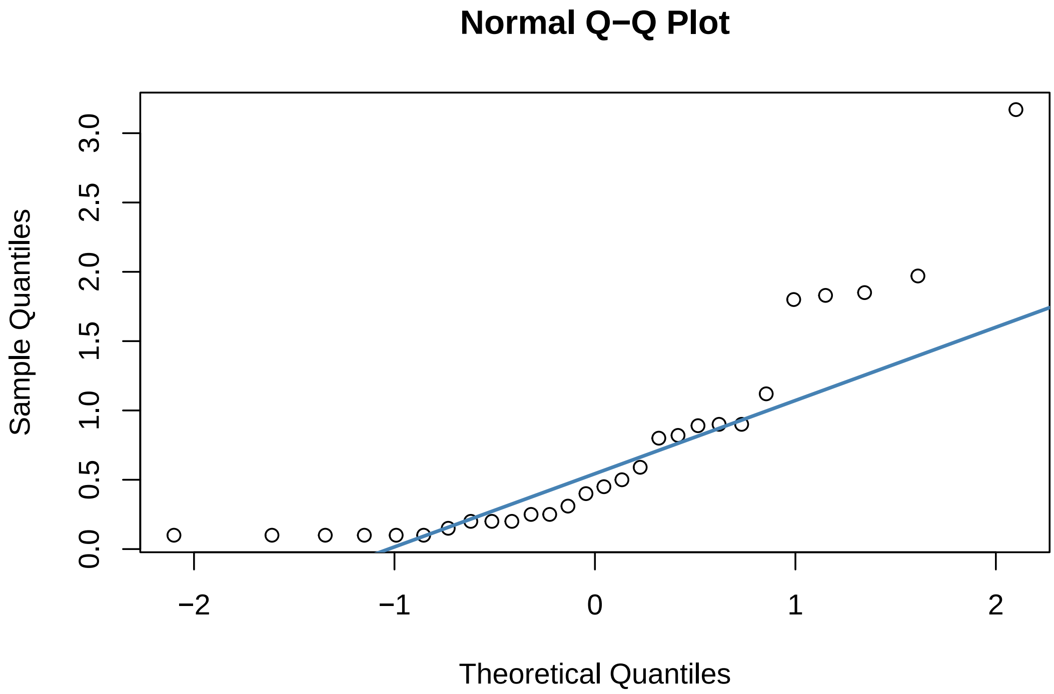
```
summary(Jul)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1000  0.2000  0.3931  0.4275  2.8000
```

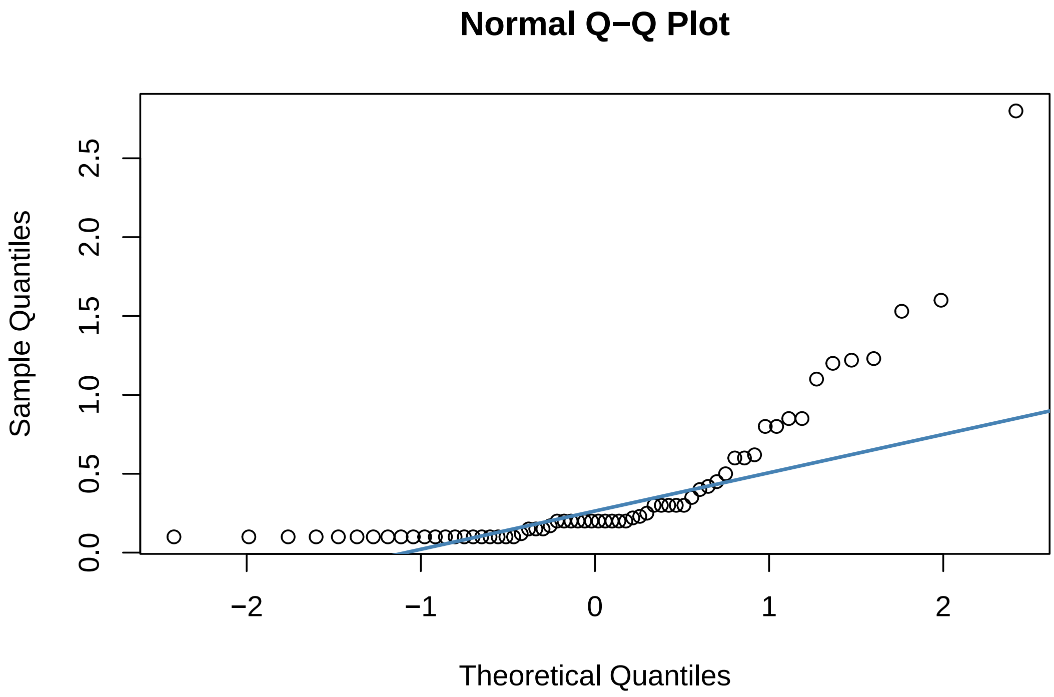
The mean and median of amount of rainfall in January is larger than July.

b Look at the QQ-plot of the data and, based on the shape, suggest what model is reasonable.

```
qqnorm(Jan, pch = 1)
qqline(Jan, col = "steelblue", lwd = 2)
```

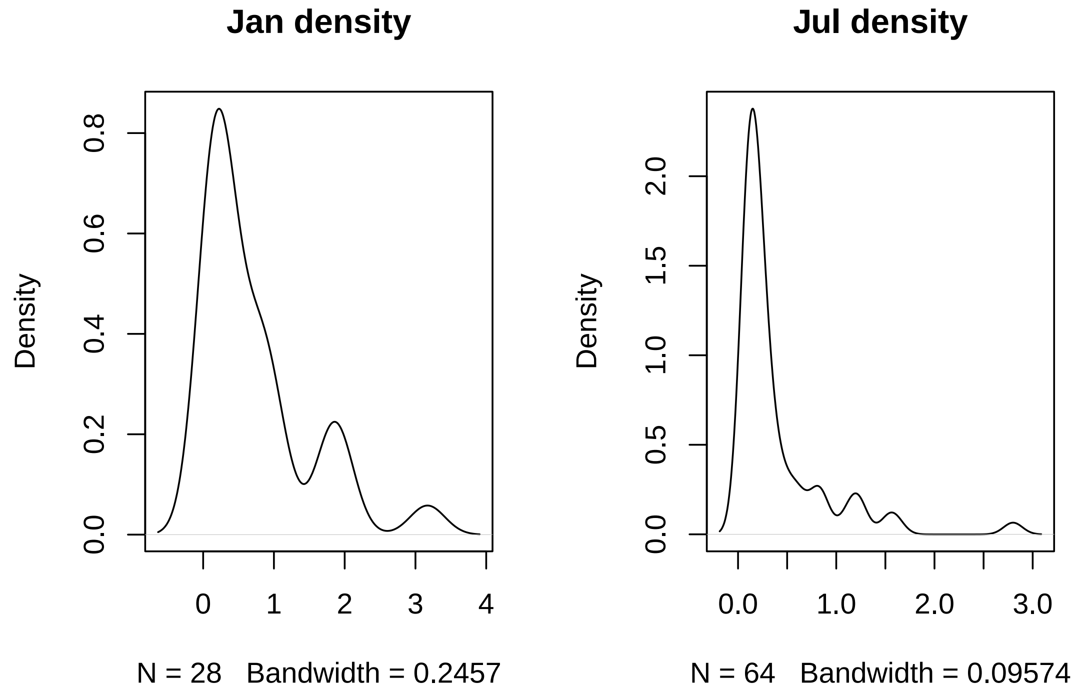


```
qqnorm(Jul, pch = 1)
qqline(Jul, col = "steelblue", lwd = 2)
```



I think from the QQ plots, we find that the sample doesn't follow normal distribution. And I also make the density plot as follow.

```
par(mfrow = c(1, 2))
plot(density(Jan),main='Jan density')
plot(density(Jul),main='Jul density')
```



From density plot, these data just look like gamma distribution. So I will try gamma distribution to fit the model.

c Fit a gamma model to the data from each month. Report the MLEs and standard errors, and draw the profile likelihoods for the mean parameters. Compare the parameters from the two months.

This is the MLEs and standard errors of January and July.

```
Jan.fit1=fitdist(Jan,'gamma','mle')
Jan.fit1
```

```
## Fitting of the distribution ’ gamma ’ by maximum likelihood
## Parameters:
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
```

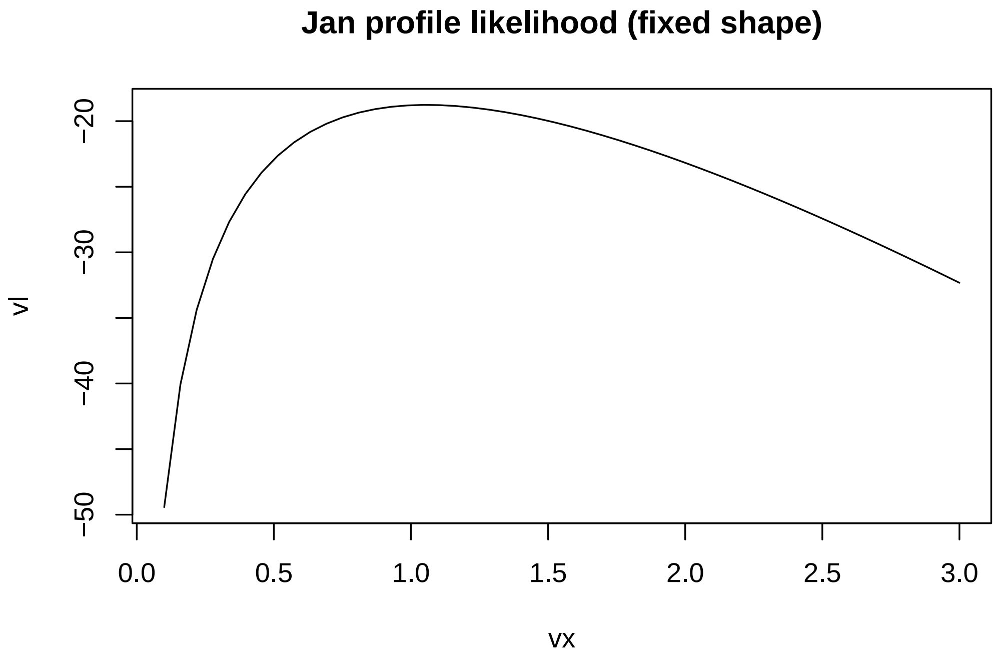
```
Jul.fit1=fitdist(Jul,'gamma','mle')
Jul.fit1
```

```
## Fitting of the distribution ’ gamma ’ by maximum likelihood
## Parameters:
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
```

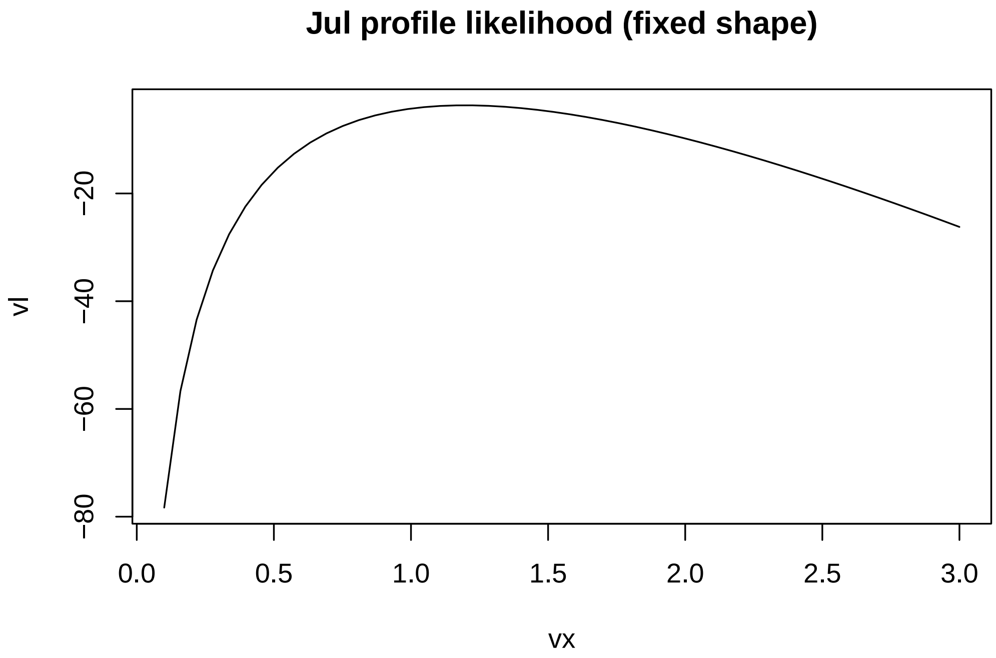
For MLE, July’s MLE is higher than January’s.

Next is the profile likelihoods for the mean parameters.
https://www.r-bloggers.com/2015/11/profile-likelihood/

```
x=Jan
prof_log_lik=function(a){
   b=(optim(1,function(z)-sum(log(dgamma(x,a,z)))))$par
   return(-sum(log(dgamma(x,a,b))))
 }
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jan profile likelihood (fixed shape)')
```

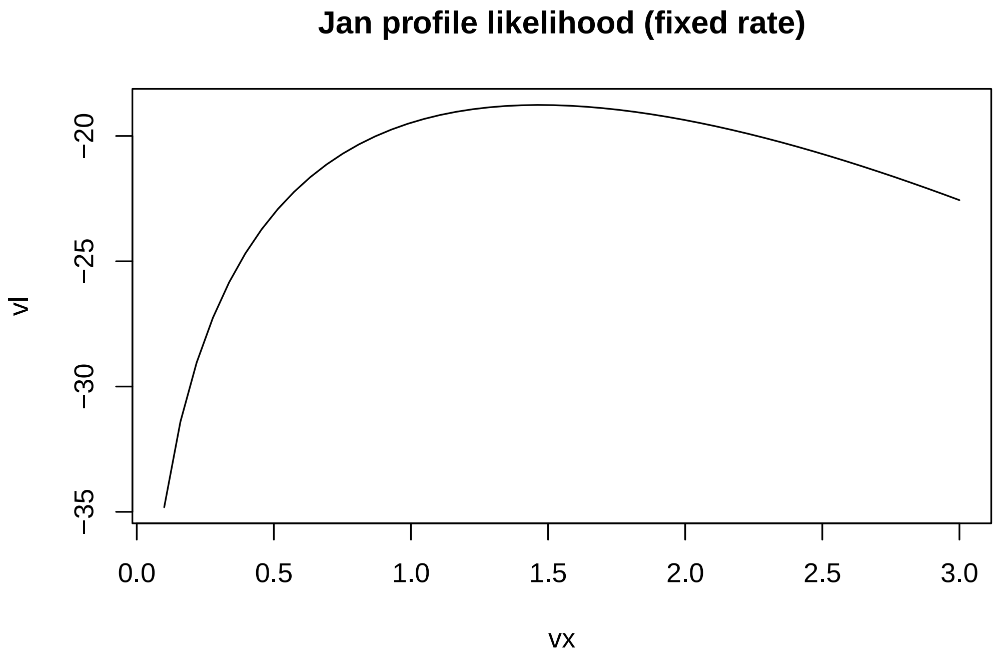


```
x=Jul
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jul profile likelihood (fixed shape)')
```

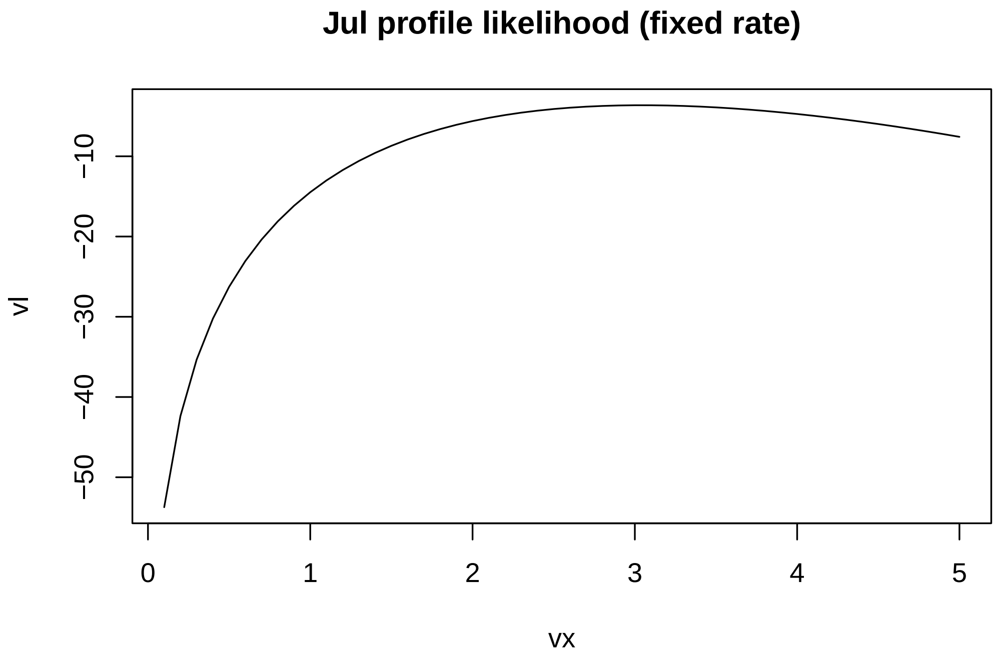


For fixed rate, we can use the same method to get the profile likelihood.

```
x=Jan
prof_log_lik=function(z){
   a=(optim(1,function(a)-sum(log(dgamma(x,a,z)))))$par
   return(-sum(log(dgamma(x,a,z))))
 }
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jan profile likelihood (fixed rate)')
```



```
x=Jul
vx=seq(.1,5,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jul profile likelihood (fixed rate)')
```

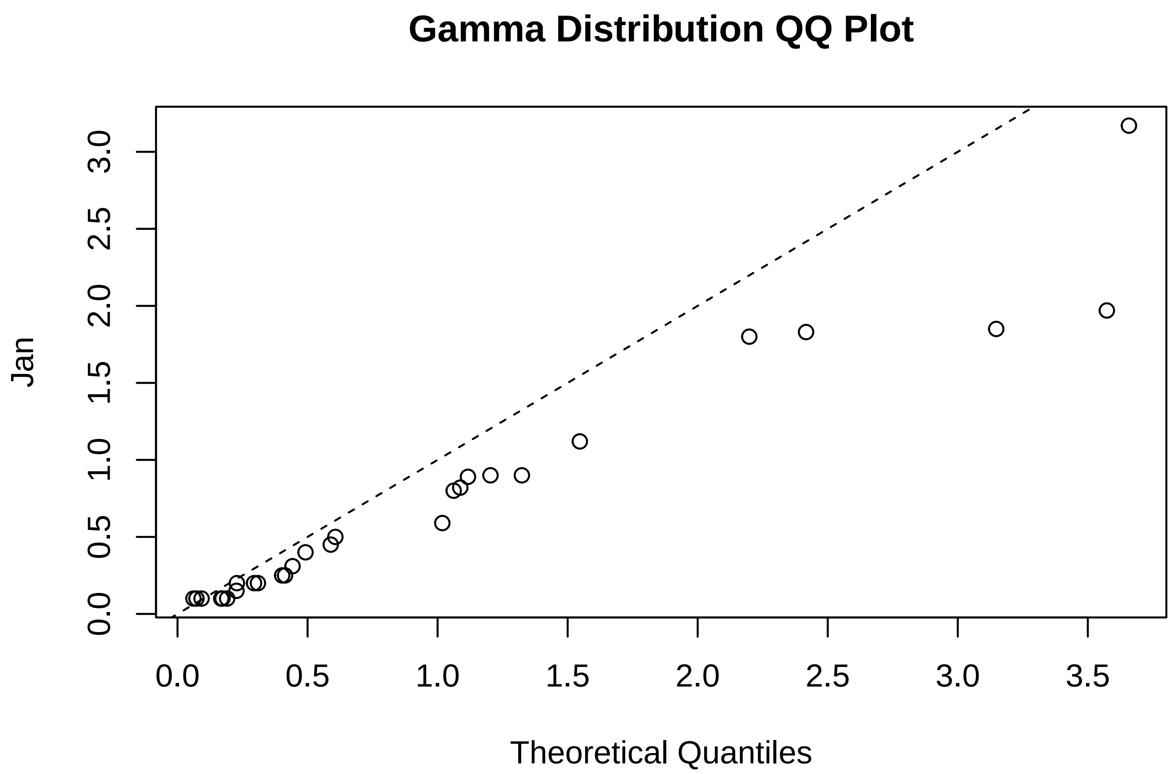


d Check the adequacy of the gamma model using a gamma QQ-plot.

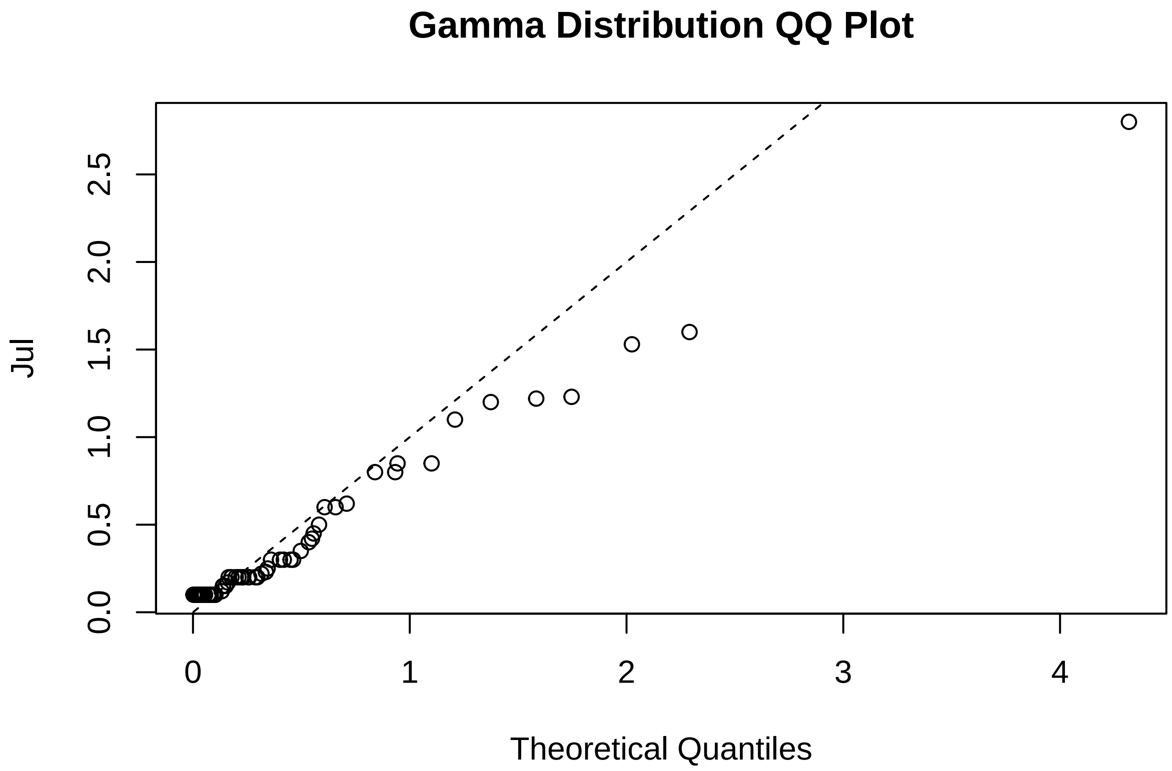
With the help of my classmate, he sent a link to me and I used the method in this link. R function-qqGamma() https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r

```
qqGamma <- function(x, ylab = deparse(substitute(x)),
                    xlab = "Theoretical Quantiles",
                    main = "Gamma Distribution QQ Plot",...){
    tx = x[!is.na(x)]
    ta = (mean(tx))^2/var(tx)
    ts = var(tx)/mean(tx)
    test = rgamma(length(tx),shape=ta,scale=ts)
    qqplot(test,tx,xlab=xlab,ylab=ylab,main=main,...)
    abline(0,1,lty=2)
}

qqGamma(Jan)
```



```
qqGamma(Jul)
```



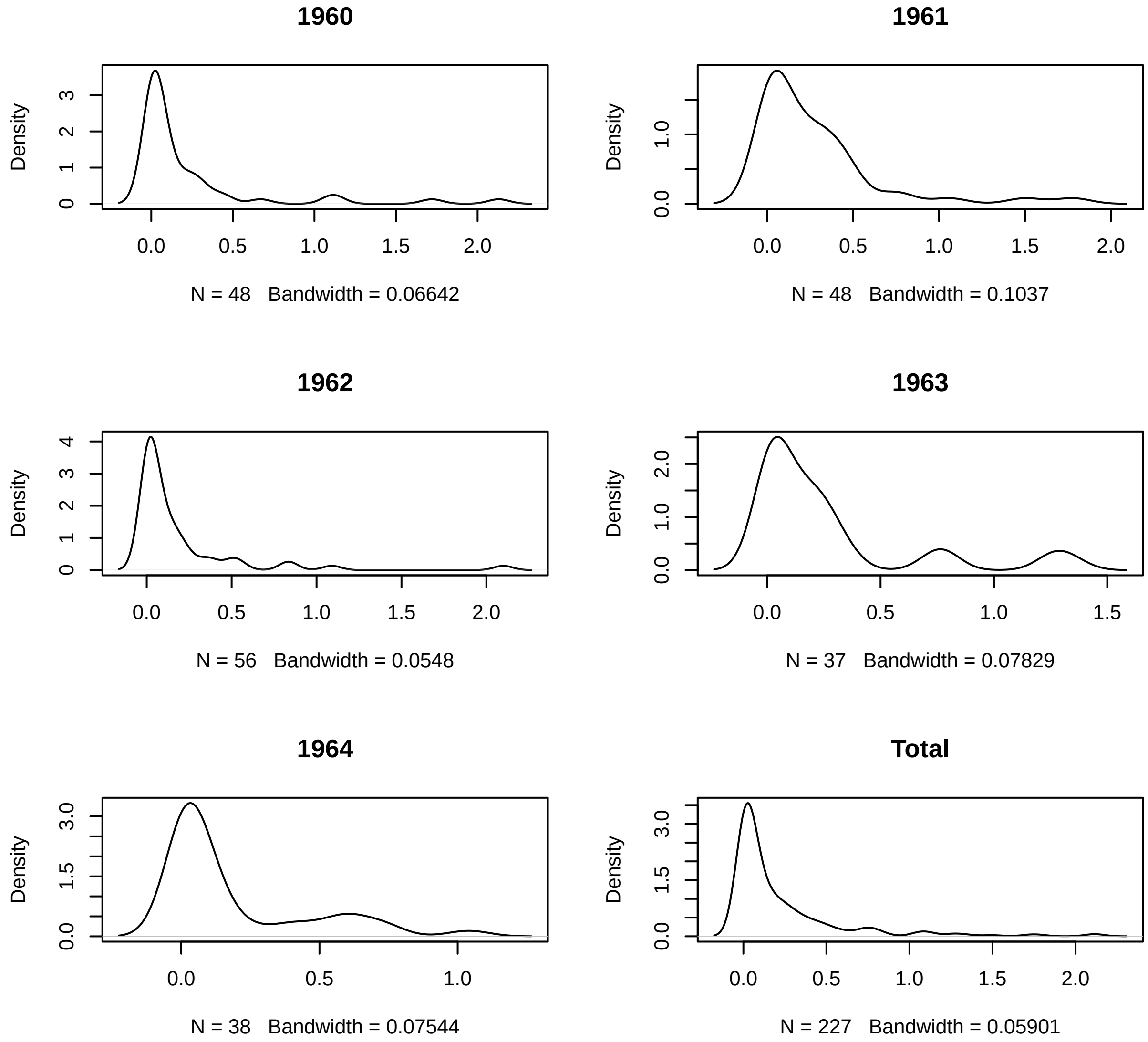
According to Gamma Q-QPlot, it seems that July is better than January.

Illinois Rain Project

Question 1

Use the data to identify the distribution of rainfall produced by the storms in southern Illinois. Estimate the parameters of the distribution using MLE. Prepare a discussion of your estimation, including how confident you are about your identification of the distribution and the accuracy of your parameter estimates.

```
rain<-read.xlsx(xlsxFile = "Illinois_rain_1960-1964.xlsx", sheet = 1, skipEmptyRows = FALSE)
par(mfrow = c(3, 2))
density(rain$`1960` %>% na.omit()) %>% plot(main='1960')
density(rain$`1961` %>% na.omit()) %>% plot(main='1961')
density(rain$`1962` %>% na.omit()) %>% plot(main='1962')
density(rain$`1963` %>% na.omit()) %>% plot(main='1963')
density(rain$`1964` %>% na.omit()) %>% plot(main='1964')
density(unlist(rain) %>%  na.omit()) %>% plot(main='Total')
```

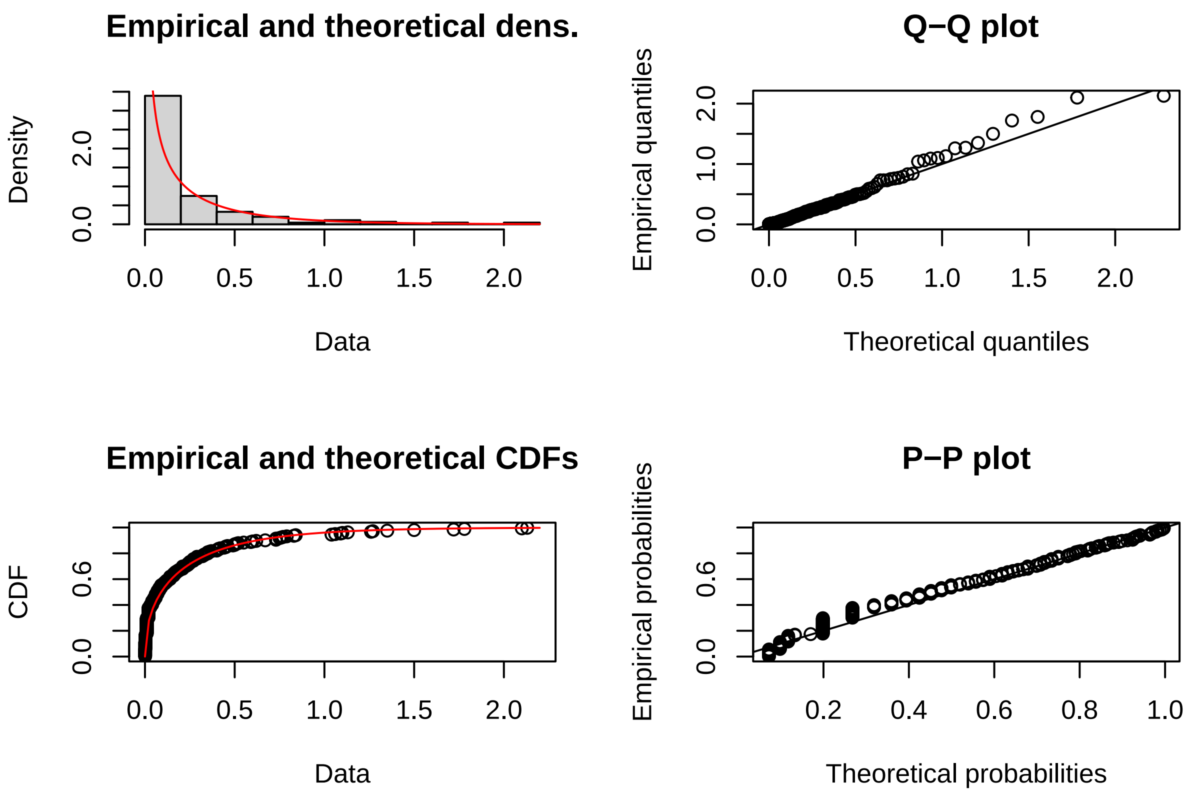


Inspired by the previous exercise, I started the whole dataset to conduct fitdist. Next I will estimate the parameters of the distribution using MLE.

```
fit<-fitdist(unlist(rain) %>%  na.omit() %>% c(),'gamma',method='mle') #MLE estimation
summary(bootdist(fit))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.442589 0.3845464 0.5183581
## rate  1.973326 1.5572294 2.5482918
```

```
plot(fit)
```



Because of plot(fit) function, we can see 4 plots. From the Empirical plots, we can see that the rad line is quite fit the histogram and points in the plot. And from Q-Q ploy and P-P plot we can see that the majority of my data points are either on or are close to the linear line. So, I'm fairly confident the distribution and the accuracy of my parameter estimates.

Question 2

Using this distribution, identify wet years and dry years. Are the wet years wet because there were more storms, because individual storms produced more rain, or for both of these reasons?

```r
rain_mean=fit$estimate[1]/fit$estimate[2]
# first get the mean of the data set
re=apply(rain,2,mean,na.rm =TRUE)
# then get mean for each year

output<-c(re,rain_mean %>% as.numeric())%>%round(4)
names(output)[6]='mean'
num_storm<-c(nrow(rain)-apply(is.na(rain),2,sum),'/')
knitr::kable(rbind(output,num_storm))  # show the results
```

	1960	1961	1962	1963	1964	mean
output	0.2203	0.2749	0.1848	0.2624	0.1871	0.2244
num_storm	48	48	56	37	38	/

We can just compare the mean of each year to the mean of these years, so we can see 1962 and 1964 are the dryer years, 1961 and 1963 are the wetter years, 1960 is the normal year. In addition more storms don't result in wet year or not. At the same time, we also found that the amount of rainfall in a single storm does not affect the wet year or not.

The same as the whole data set, I also conduct fitdist on each year. The mean of whole table is quite the same as the whole data set.

Question 3

To what extent do you believe the results of your analysis are generalization? What do you think the next steps would be after the analysis?

During my practice, I though that the data was too small to be verification from my opinion. At the same time, I also though that because the total number of years is too small, there was an error in the average precipitation calculated.

Therefore, I think the next step should be to collect many more years of data, or consider data from various regions at the same time, so as to expand the data set and get more accurate results.