# MA677 final

zby

5/13/2022

## 4.25

```
pdf <- function(x, a=0, b=1) dunif(x, a,b) #pdf function
cdf <- function(x, a=0, b=1) punif(x, a,b, lower.tail=FALSE) #cdf function
integrand <- function(x,r,n) {
  x * (1 - cdf(x))^(r-1) * cdf(x)^(n-r) * pdf(x)
}
E <- function(r,n) {
  (1/beta(r,n-r+1)) * integrate(integrand,-Inf,Inf, r, n)$value
}
medianprrox<-function(k,n){
  m<-(k-1/3)/(n+1/3)
  return(m)
}
E(2.5,5)
```

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

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

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

```
E(5,10)
```

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

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

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

The results are similar.

## 4.27

```r
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)**

```r
summary(Jan)
```

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

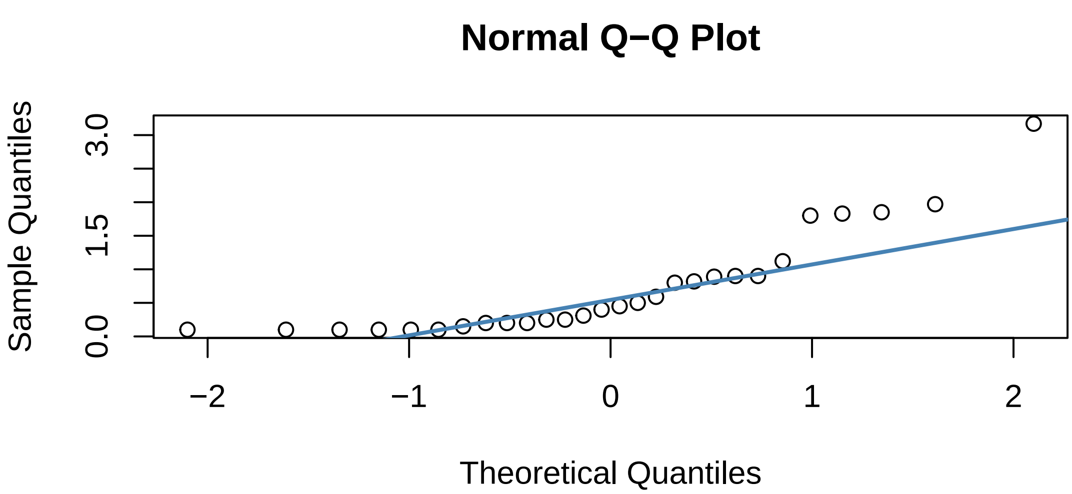
```r
summary(Jul)
```

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

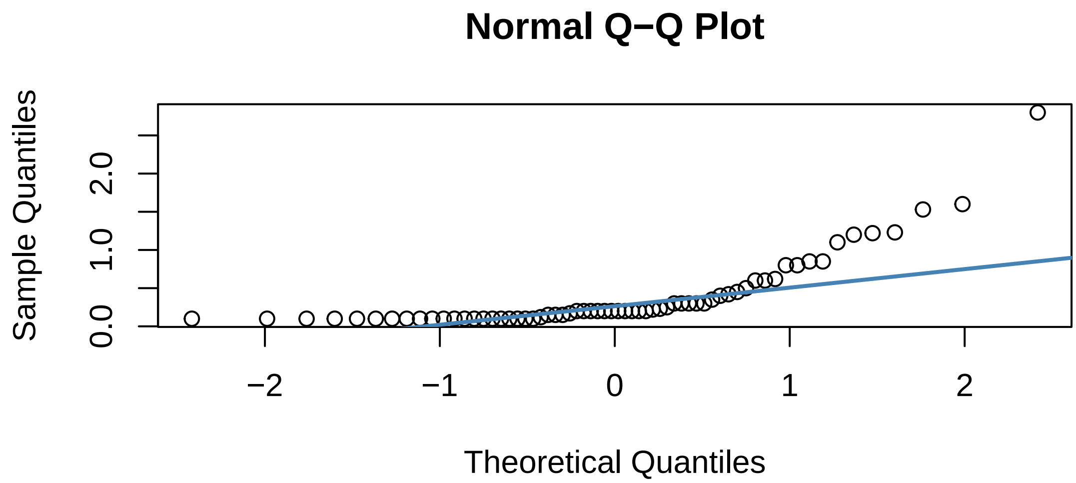
We can see that 1st, Median, Mean 3rd Max of Jan are higher than those of Jul.

**(b)**

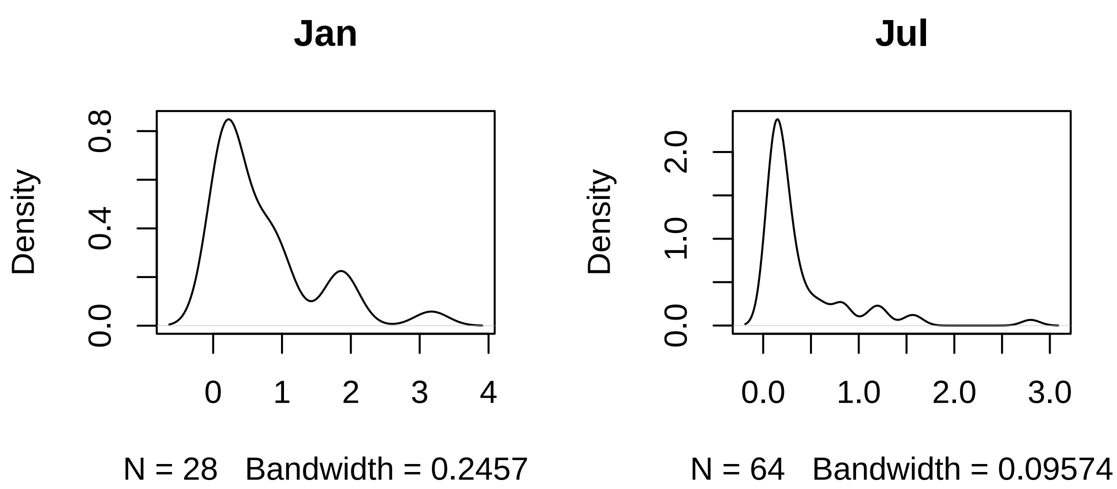
```r
# Reference:https://towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-9fbff

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



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




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




From the above plot, we can conclude that it is gamma distribution.

### (c)

I used MLE method to solve this problem.

```
Jan.fit1=fitdist(Jan,'gamma','mle')
Jul.fit1=fitdist(Jul,'gamma','mle')
exp(Jan.fit1$loglik)
```

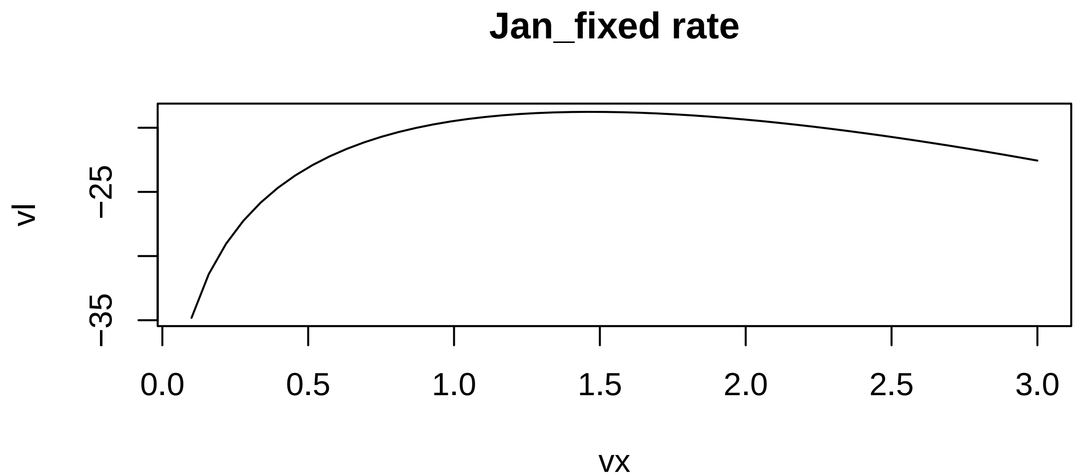
```
## [1] 7.11117e-09
```

```
exp(Jul.fit1$loglik)
```

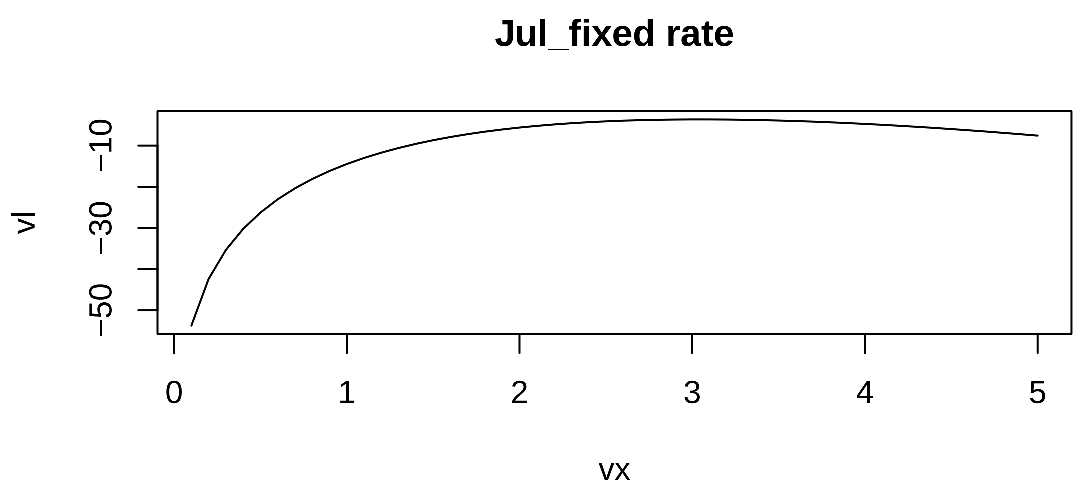
```
## [1] 0.02638693
```

We can use the same method to get the profile likelihood of fixed rate.

```
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_fixed rate')
```

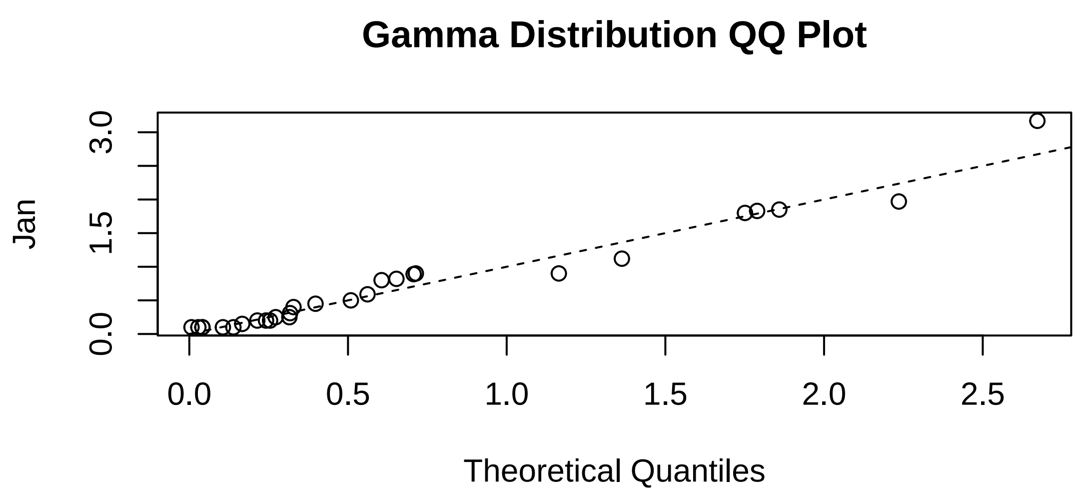



```
x=Jul

vx=seq(.1,5,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jul_fixed rate')
```

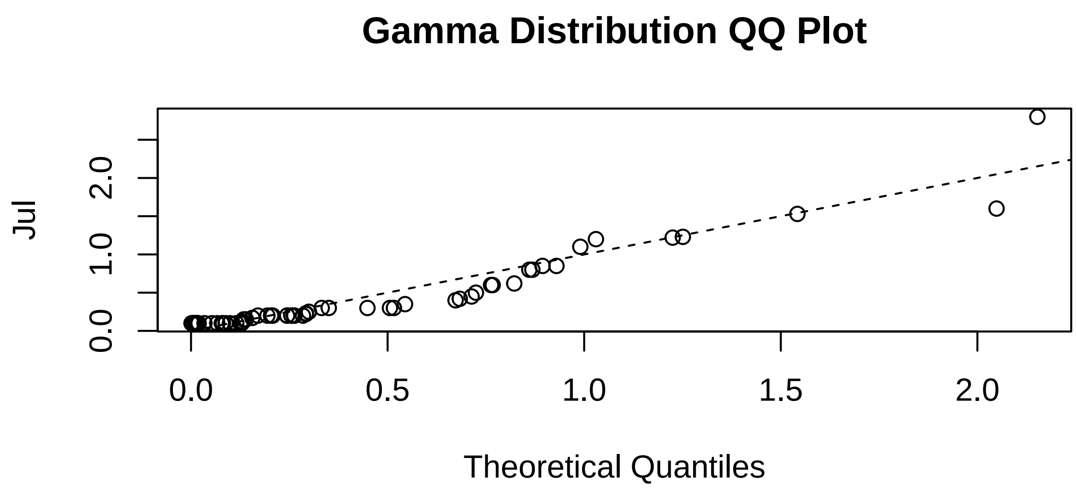


**(d)**

```r
# reference:qpToolkit
# 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",...)
{
    # Plot qq-plot for gamma distributed variable

    xx = x[!is.na(x)]
    aa = (mean(xx))^2 / var(xx)
    ss = var(xx) / mean(xx)
    test = rgamma(length(xx), shape = aa, scale = ss)

    qqplot(test, xx, xlab = xlab, ylab = ylab, main = main,...)
    abline(0,1, lty = 2)
}


qqGamma(Jan)
```

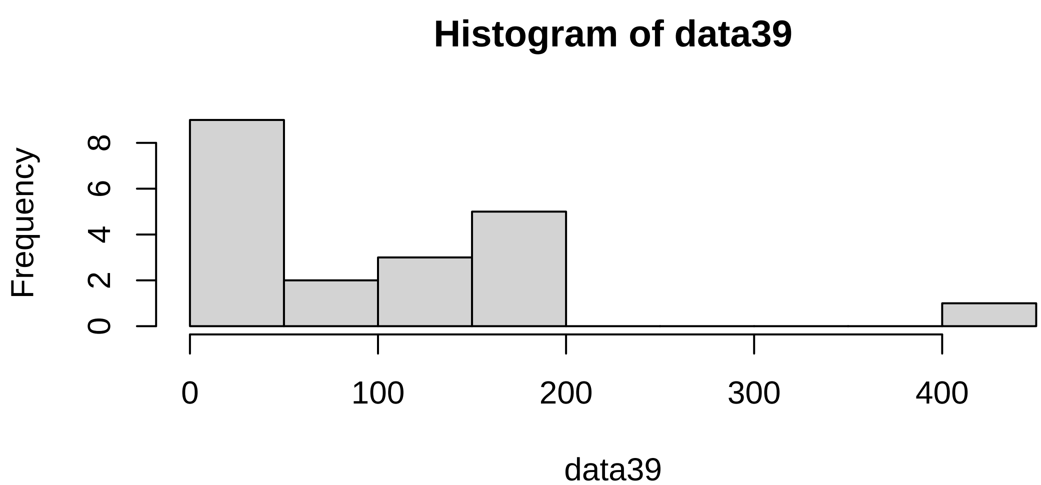



```r
qqGamma(Jul)
```

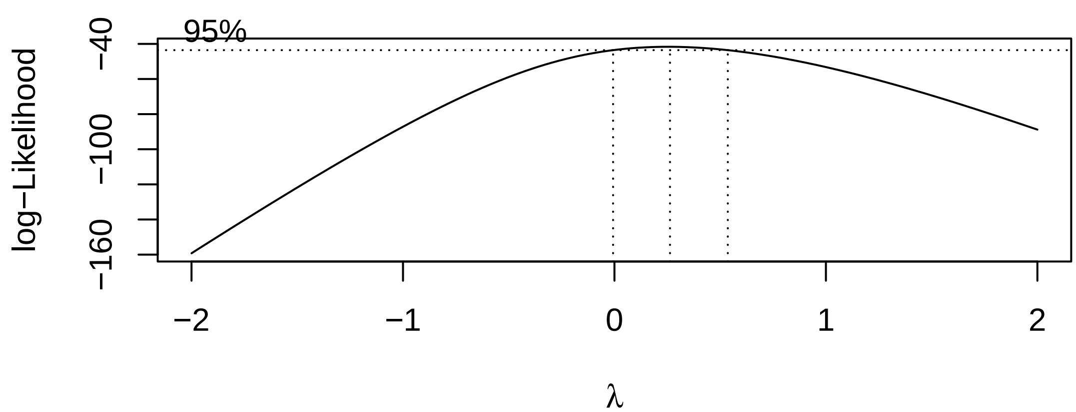



**4.39**

```
data39<-c(0.4,1.0,1.9,3.0,5.5,8.1,12.1,25.6,50.0,56.0,70.0,115.0,115.0,119.5,154.5,157.0,175.0,179.0,18
hist(data39)
```



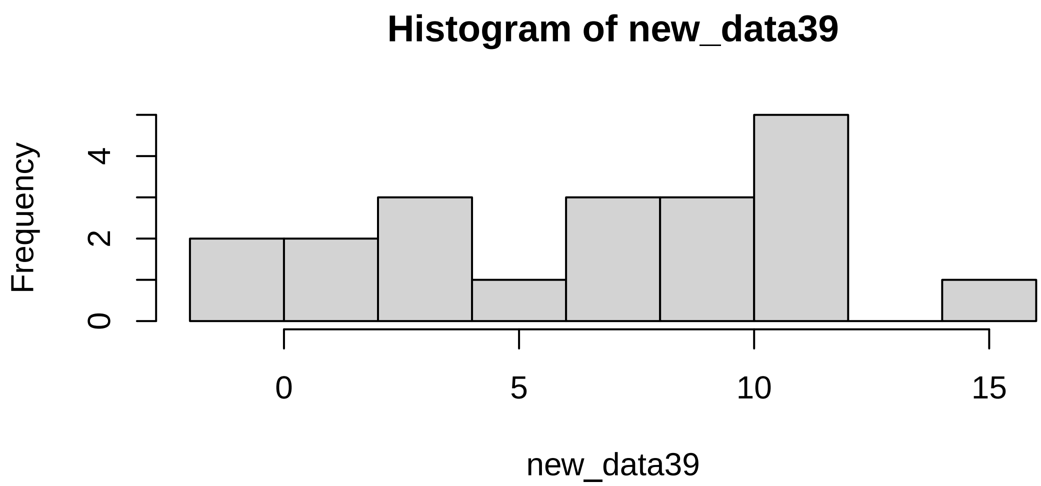

```
b <- boxcox(lm(data39 ~ 1))
```



```
lambda <- b$x[which.max(b$y)]
lambda #lambda=0.2626263
```

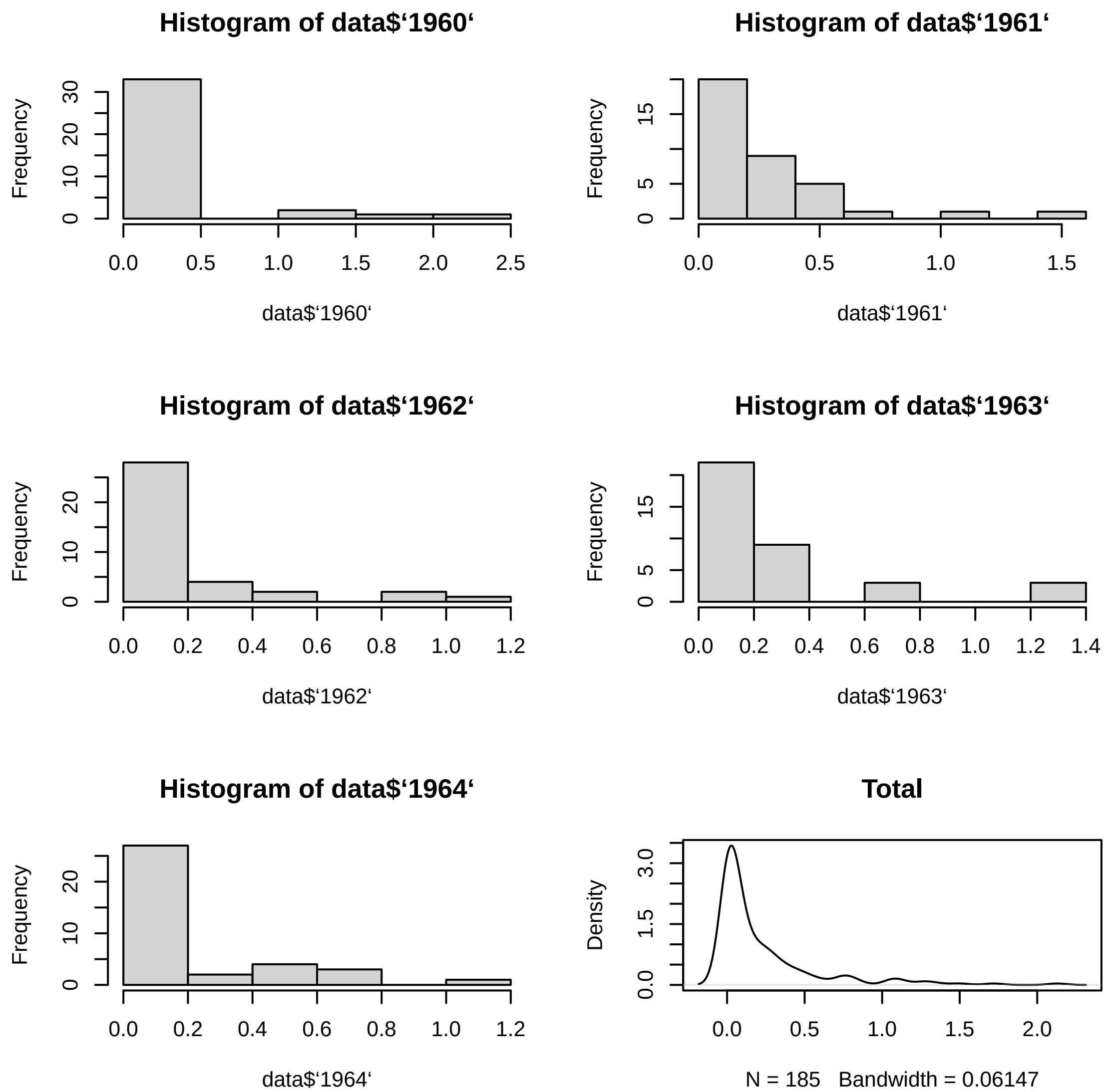
```
## [1] 0.2626263
```

```
new_data39 <- (data39 ^ lambda - 1) / lambda
hist(new_data39)
```




### In All Likelihood

```
data<-read.xlsx("/Users/zhangbiyao/Desktop/Illinois_rain_1960-1964.xlsx")
data<-na.omit(data)
#summary(data)
#is.na(data) # returns a vector
par(mfrow = c(3, 2))
hist(data$`1960`)
hist(data$`1961`)
hist(data$`1962`)
hist(data$`1963`)
hist(data$`1964`)
density(unlist(data)) %>% plot(main='Total')
```

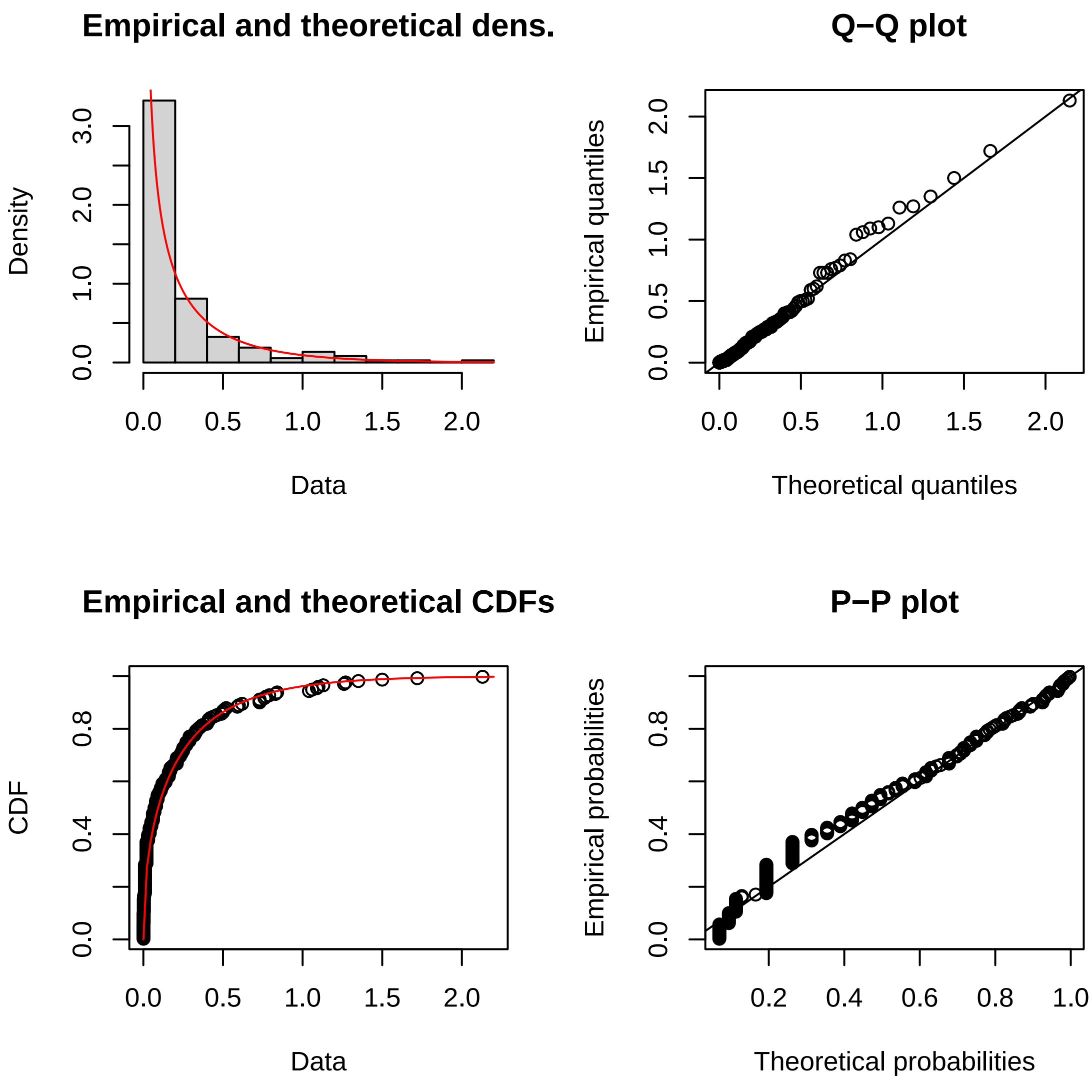



Firstly, I tried to conduct fitdist.

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

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4537441 0.3847742 0.5466761
## rate  2.0436250 1.5640144 2.7591199
```

```
plot(fit)
```




Secondly, I used this distribution, identify wet years and dry years.

```
rain_mean=fit$estimate[1]/fit$estimate[2] #get mean for whole dataset
re=apply(data,2,mean,na.rm =TRUE) # get mean for each year
out<-c(re,rain_mean %>% as.numeric() %>% round(4))
names(out)[6]='mean'
num_storm<-c(nrow(data)-apply(is.na(data),2,sum),'/')
knitr::kable(rbind(out,num_storm))
```

| | 1960 | 1961 | 1962 | 1963 | 1964 | mean |
|---|---|---|---|---|---|---|
| out | 0.245864864864865 | 0.253972972972973 | 0.163729729729730 | 0.262432432432432 | 0.190810810810811 | 0.2234 |
| num_storm | 37 | 37 | 37 | 37 | 37 | / |

1962, 1964 are dryer years, 1961 and 1963 are wetter years. From the results, we can conclude that storms don't result in wet year and more rain in individual storm don't y result in wet year.

I think the nest step is to confirm the results, because the dataset is not enough. Huff didn't have relaible data, so he can't do deep research.

## What I learned?

I learned important statistical methods (e.g MLE, MSE, Empirical Likelihood), and I learned R to implement these methods. I am interested in empirical method, I want to read related textbooks.