# 677 final

Qiannan Shen

5/11/2022

## 4.25

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

# distribution of the order statistics
integrand <- function(x, r, n){
  x * (1 - F(x))^(r-1) * F(x)^(n-r) * f(x)
}

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

# approximation
median <- function(k, n){
  m <- (k-1/3)/(n+1/3)
  return(m)
}

E(2.5, 5)
```

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

```
median(2.5, 5)
```

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

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

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

```
median(5,10)
```

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

## 4.27

```
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)
July <- 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,
                     0.20,0.10,0.10,0.10,0.42,0.85,1.60,0.10,0.25,0.10,0.20,0.10)
```

### a

```
summary(Jan)
```

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

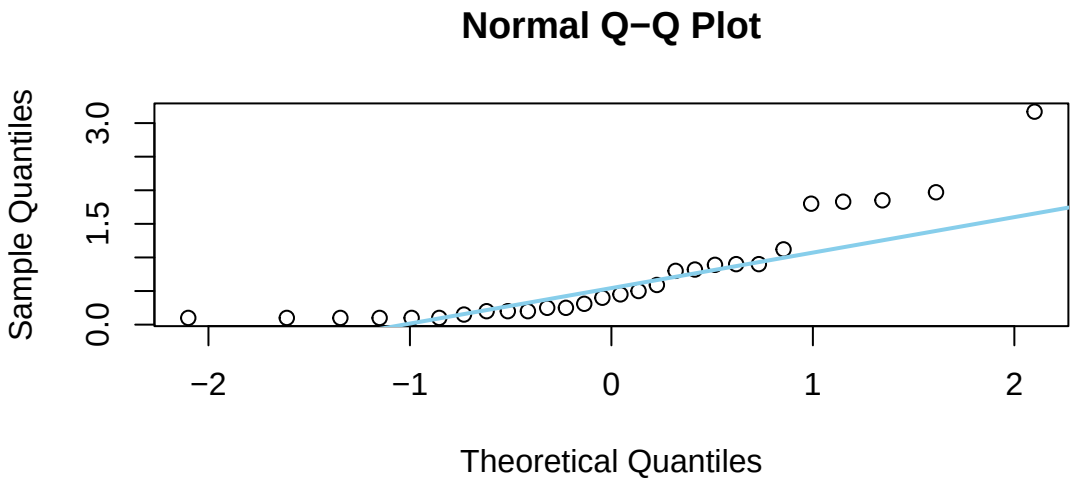
```
summary(July)
```

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

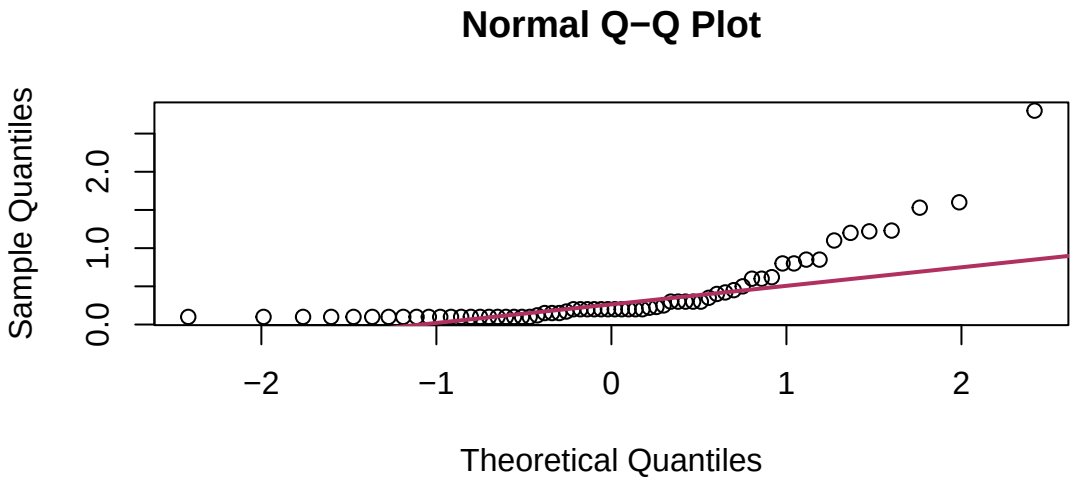
Except the minimum, the other values we got in the summary of January 1940 are higher than that of July 1940.

### b

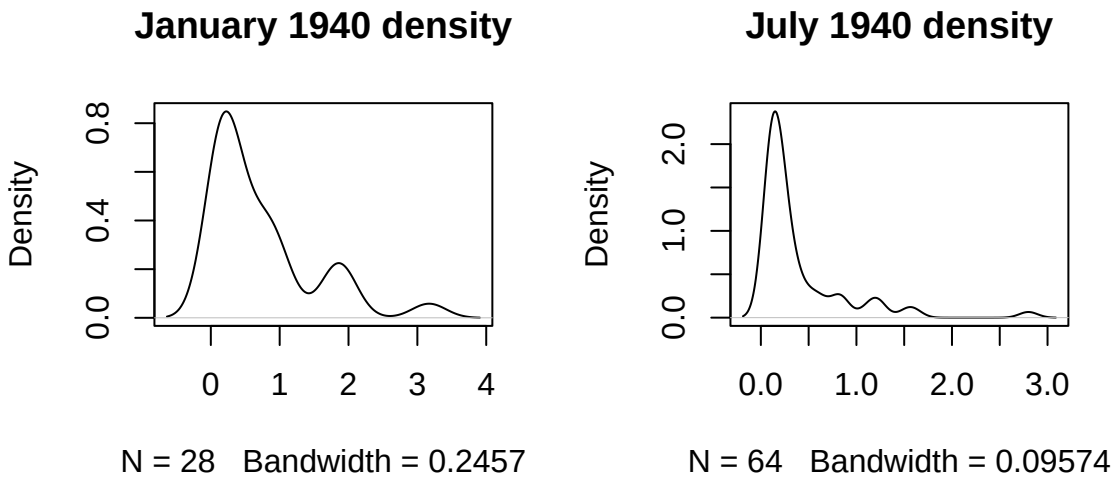
```
qqnorm(Jan, pch = 1)
qqline(Jan, col = 'skyblue', lwd = 2)
```



```
qqnorm(July, pch = 1)
qqline(July, col = 'maroon', lwd = 2)
```




```
par(mfrow = c(1,2))
plot(density(Jan), main = 'January 1940 density')
plot(density(July), main = 'July 1940 density')
```




According the qq-plot, the data doesn't follow a normal distribution. After ploting the density plot, the data seems to follow a gamma distribution.

**c**

```
library(fitdistrplus)
fit1 <- fitdist(Jan, distr = 'gamma', method = 'mle')
fit7 <- fitdist(July, distr = 'gamma', method = 'mle')
summary(fit1)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
## Loglikelihood:  -18.7616   AIC:  41.5232   BIC:  44.18761
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.7893943
## rate  0.7893943 1.0000000
```

```
summary(fit7)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
## Loglikelihood:  -3.634886   AIC:  11.26977   BIC:  15.58754
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.8103948
## rate  0.8103948 1.0000000
```

```
#MLE
exp(fit1$loglik)
```

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

```
exp(fit7$loglik)
```

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

From MLE, Jul's MLE is higher than the one of Jan. This means that Jul's model is better than Jan's.

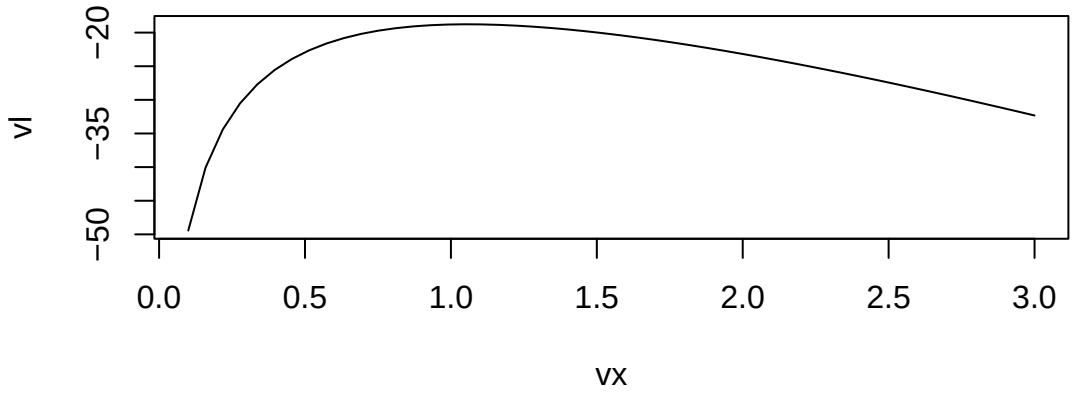
```
#sd
fit1$sd
```

```
##     shape      rate
## 0.2497495 0.4396202
```

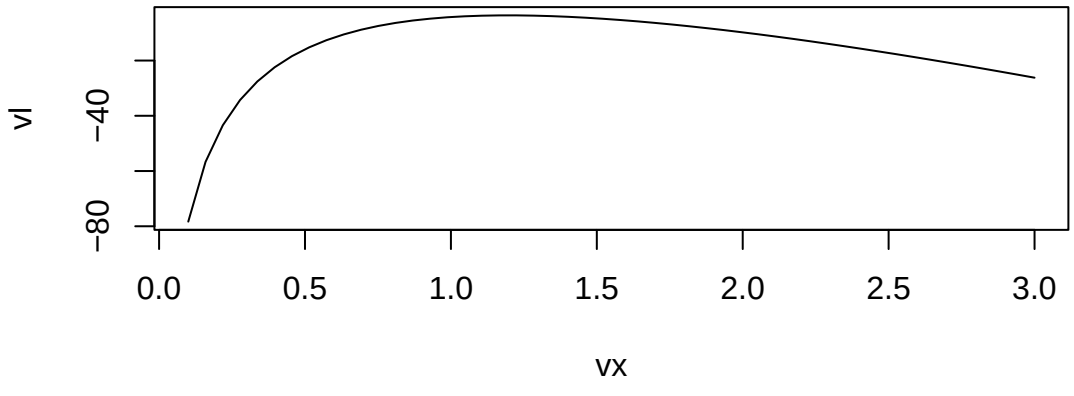
```
fit7$sd
```

```
##     shape      rate
## 0.1891196 0.5936302
```

```r
#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")
```

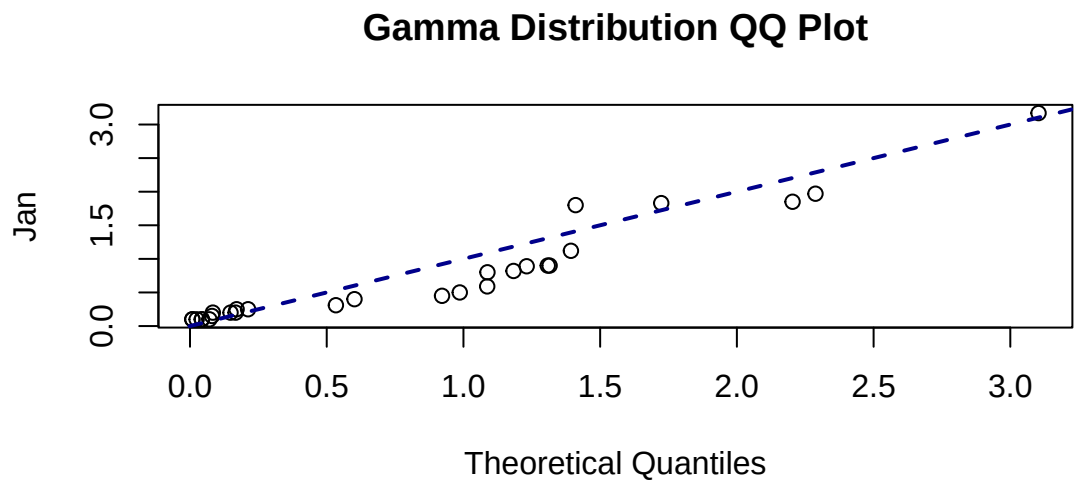



```r
x=July
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l")
```

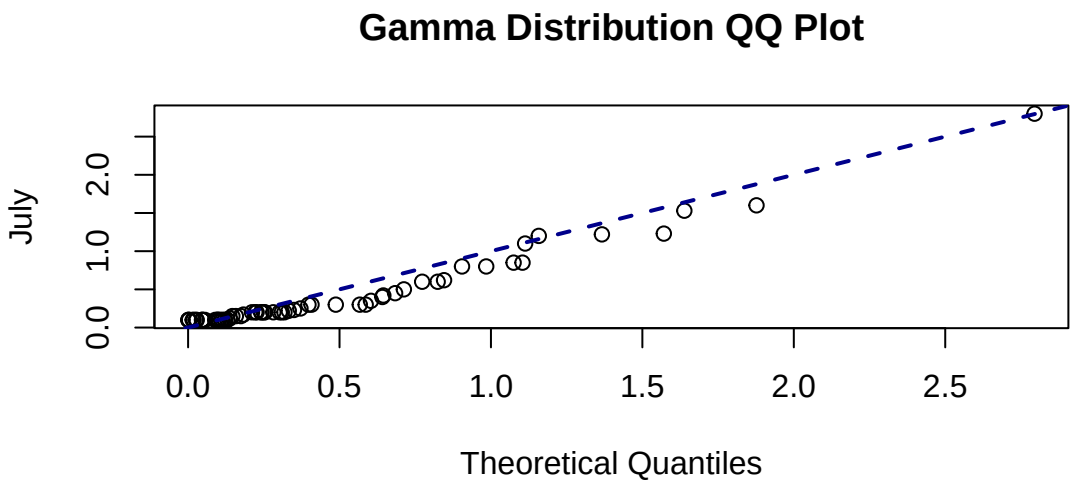


**d**

```
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, col = "darkblue", lwd = 2)
}

qqGamma(Jan)
```



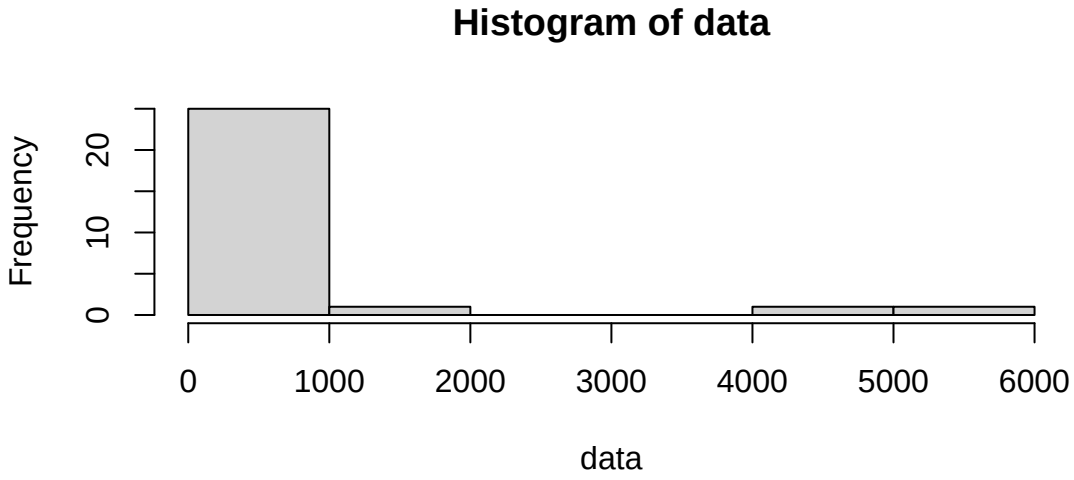

```
qqGamma(July)
```



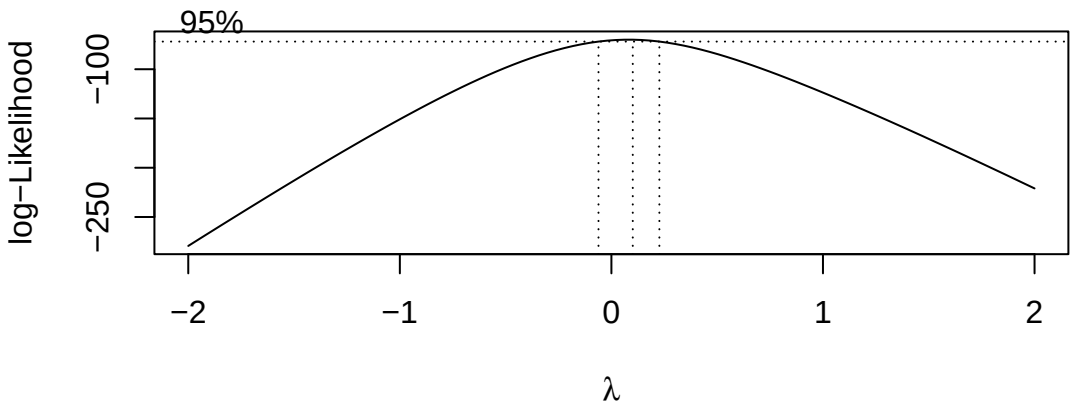

After comparaing the two gamma qq-plot, july 1940 should be better.

## 4.39

```
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,
          419.0, 423.0, 440.0, 655.0, 680.0, 1320.0, 4603.0, 5712.0)
hist(data)
```



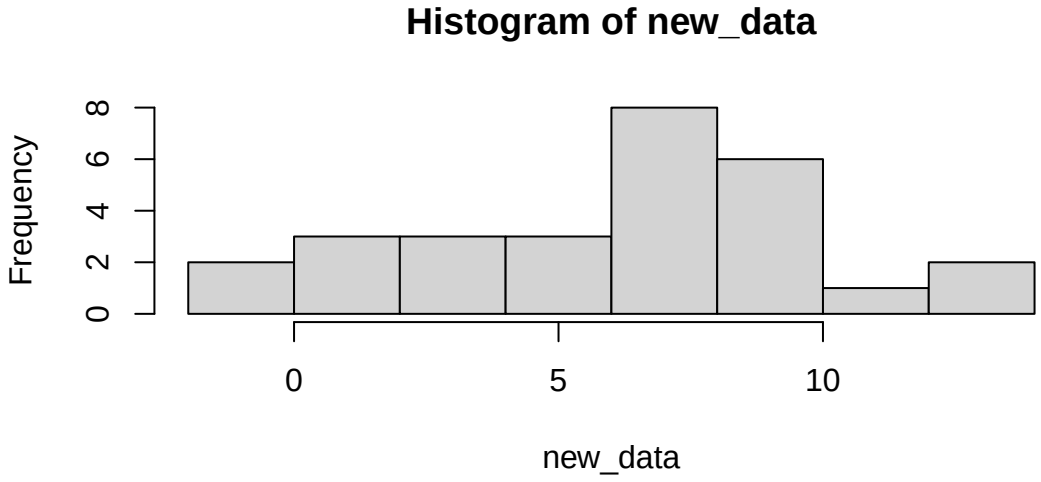

```
# Conduct boxcox transformation
b <- boxcox(lm(data ~ 1))
```




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

```
## [1] 0.1010101
```

```
new_data <- (data ^ lambda - 1) / lambda
hist(new_data)
```



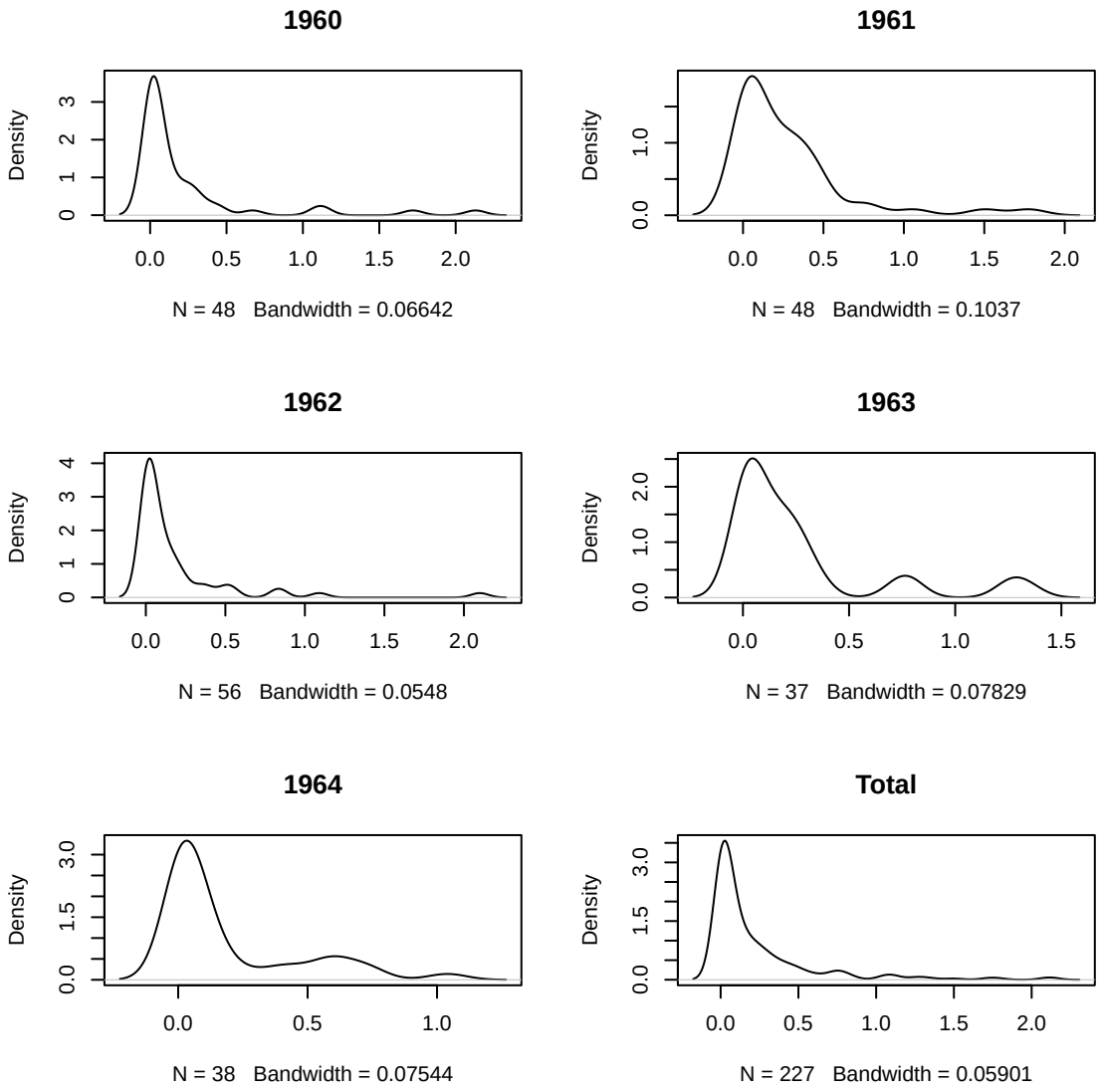
# Illinois

## a

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.

First, I draw the density plot ffor the rainfall from 1960 to 1964 and decide the distribution we would use in our model. Shown as below, using gamma distribution is a good choice according to the density shape.

```
rain=read.xlsx('Illinois_rain_1960-1964.xlsx')
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')
```




```
fit1<-fitdist(unlist(rain) %>%  na.omit() %>% c(),'gamma',method='mle') #MLE estimation
fit2<-fitdist(unlist(rain) %>%  na.omit() %>% c(),'gamma',method='mse') #MSE estimation
```

```
boot_mle <- bootdist(fit1)
summary(boot_mle) #boot get confidence interval
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4434727 0.3818621 0.5151849
## rate  1.9987090 1.5829253 2.5872403
```

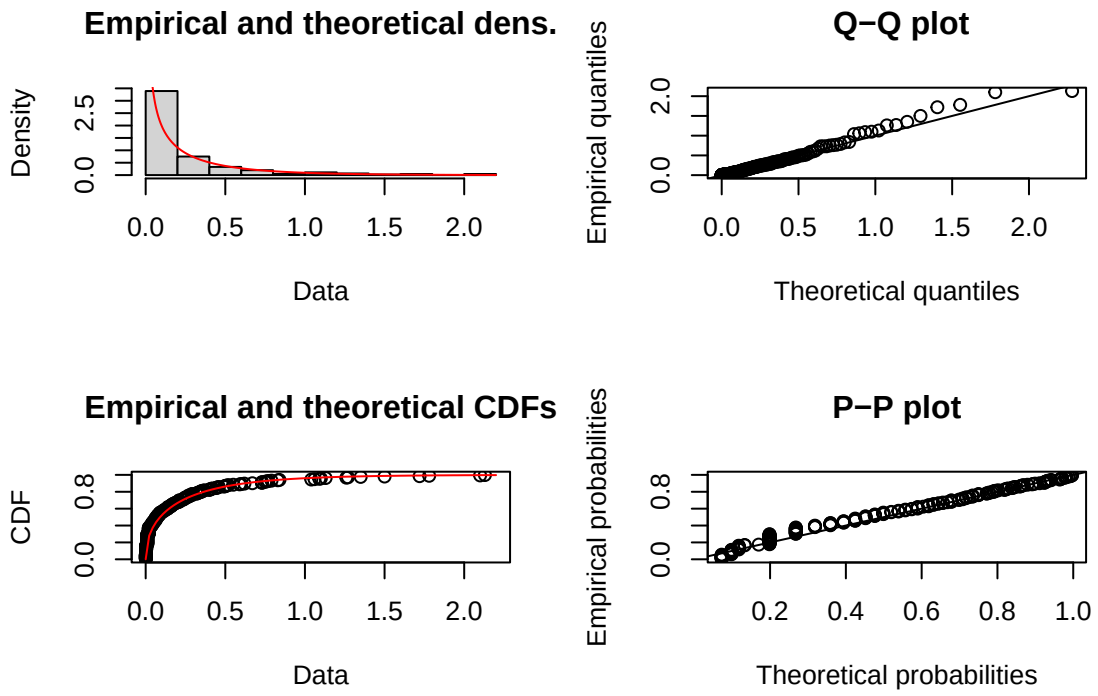
```
boot_mse <- bootdist(fit2)
summary(boot_mse) #boot get confidence interval
```

```
## Parametric bootstrap medians and 95% percentile CI
```

```
##          Median      2.5%     97.5%
## shape 0.7199484 0.6155877 0.8442792
## rate  1.3420655 1.0927339 1.6925691
```

According to the summary for mle and mse methods, we could see that the confidence interval for mle method is wider than that for mse method, which means the estimation is more reliable. Therefore, it is better to choose the MLE method. As we can the graphs below, almost all the points lies on the line in the qq-plot, empirical cdf plot and pp-plot.

```
plot(fit1)
```




## b

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?

First, I calculate and use the average rainfall in the five years as a indicator indentifying the wet or dry years, and count the number of storms for each year to find the reason.

```
rain_mean=fit1$estimate[1]/fit1$estimate[2] #get mean for whole dataset
rain_mean
```

```
##     shape
## 0.2243635
```

```r
re=apply(rain,2,mean,na.rm =TRUE) # get mean for each year

rain_year<-c(re,rain_mean %>% as.numeric() %>% round(4))
names(rain_year)[6]='mean'
rain_year
```

```
##      1960      1961      1962      1963      1964      mean
## 0.2202917 0.2749375 0.1847500 0.2624324 0.1871053 0.2244000
```

```r
num_storm<-c(nrow(rain)-apply(is.na(rain),2,sum),'/')
num_storm
```

```
## 1960 1961 1962 1963 1964
## "48" "48" "56" "37" "38"  "/"
```

Use table to show the result:

| Year | 1960 | 1961 | 1962 | 1963 | 1964 | 5-year average |
|---|---|---|---|---|---|---|
| Average | 0.22029 | 0.27494 | 0.18475 | 0.26243 | 0.18711 | 0.22440 |
| Num storm | 48 | 48 | 56 | 37 | 38 | 45.4 |

Compared to average storm rainfall each year with the average for five years, we could include that the year 1962 and 1964 were dry years, the year 1960, 1961 and 1963 were wet years. However, when we compare the number of storms for each, we can conclude that the year 1960, 1961 and 1962 was a wet year. As a result, we may conclude that more storms don't necessarily result in wet year and more average rainfall in individual storm don't necessarily result in wet year, both of these reasons have influence on whether a year is a wet or dry year.

**c**

To what extent do you believe the results of your analysis are generalizable? From my perspective, 5-year data is not enough for my analysis to be generalizable. From Floyd Huff's Time Distribution Rainall in Heavy Storms in 1967. He used Network data for the 11-year period 1955 - 1966.

From the perspective of the report by Floyd Huff, he is more concentrated to the theoretical description, he did not build a reliable model based on his thesis. In my perspective, in order to have a better prediction, only having 5-years' data is not enough and we need to collect more data to get a more reliable model. Also, doing more detailed research could help.

# Citation:


1. I consulted this final project from Yuli Jin. From Yuli Jin, I learnt to draw the profile likelihood and use bootstrap to generate confidence interval for gamma distribution.
2. QQ Gamma Plot: https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r