677 Final

Yongrong Chai

5/12/2022

In All Likelihood

4.25

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

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


# Expectation
Exp <- function(r,n) {
  (1/beta(r,n-r+1)) * integrate(order_statistics,-Inf,Inf, r, n)$value
}

# Approximation function
approx<-function(k,n){
  return((k-1/3)/(n+1/3))
}
# for n=5
Exp(2.5,5)
```

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

```
approx(2.5,5)
```

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

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

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

```
approx(5,10)
```

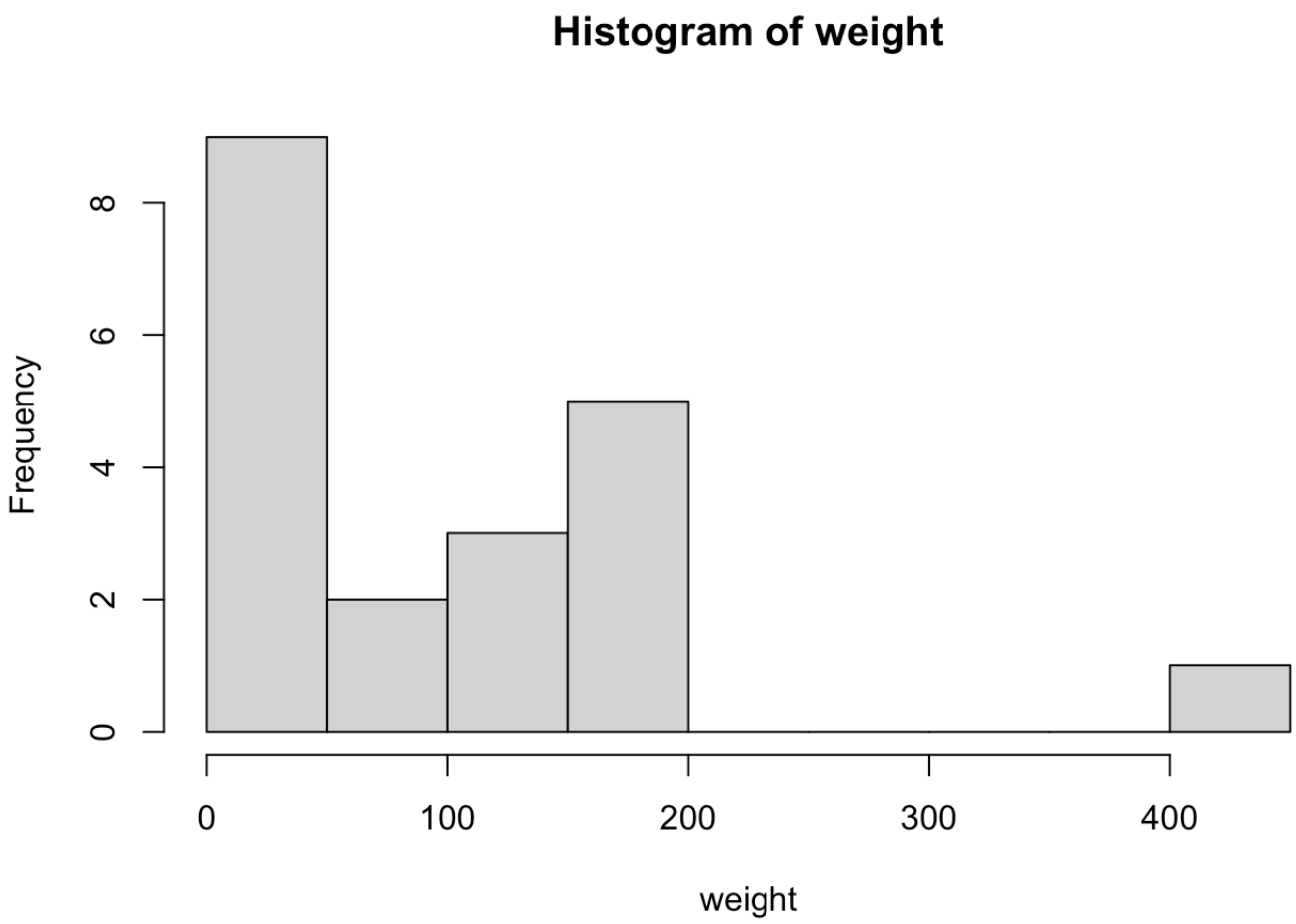
```
## [1] 0.4516129
```

Based on the result, they are really close.

4.39

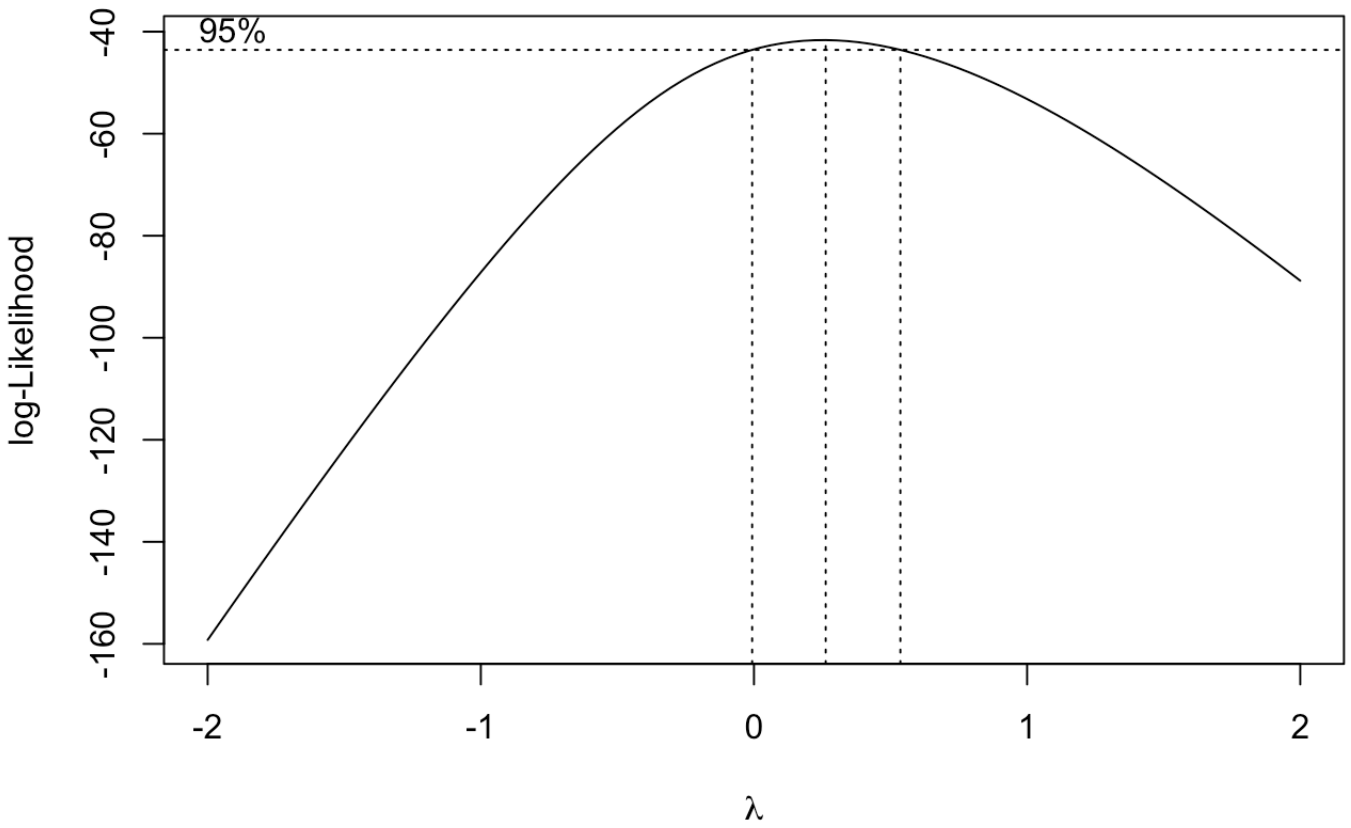
Here is the data for 28 species of animals

```
weight <- 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(weight)
```



Boxcox transformation

```
library(MASS)
b_trans <- boxcox(lm(weight ~ 1))
```

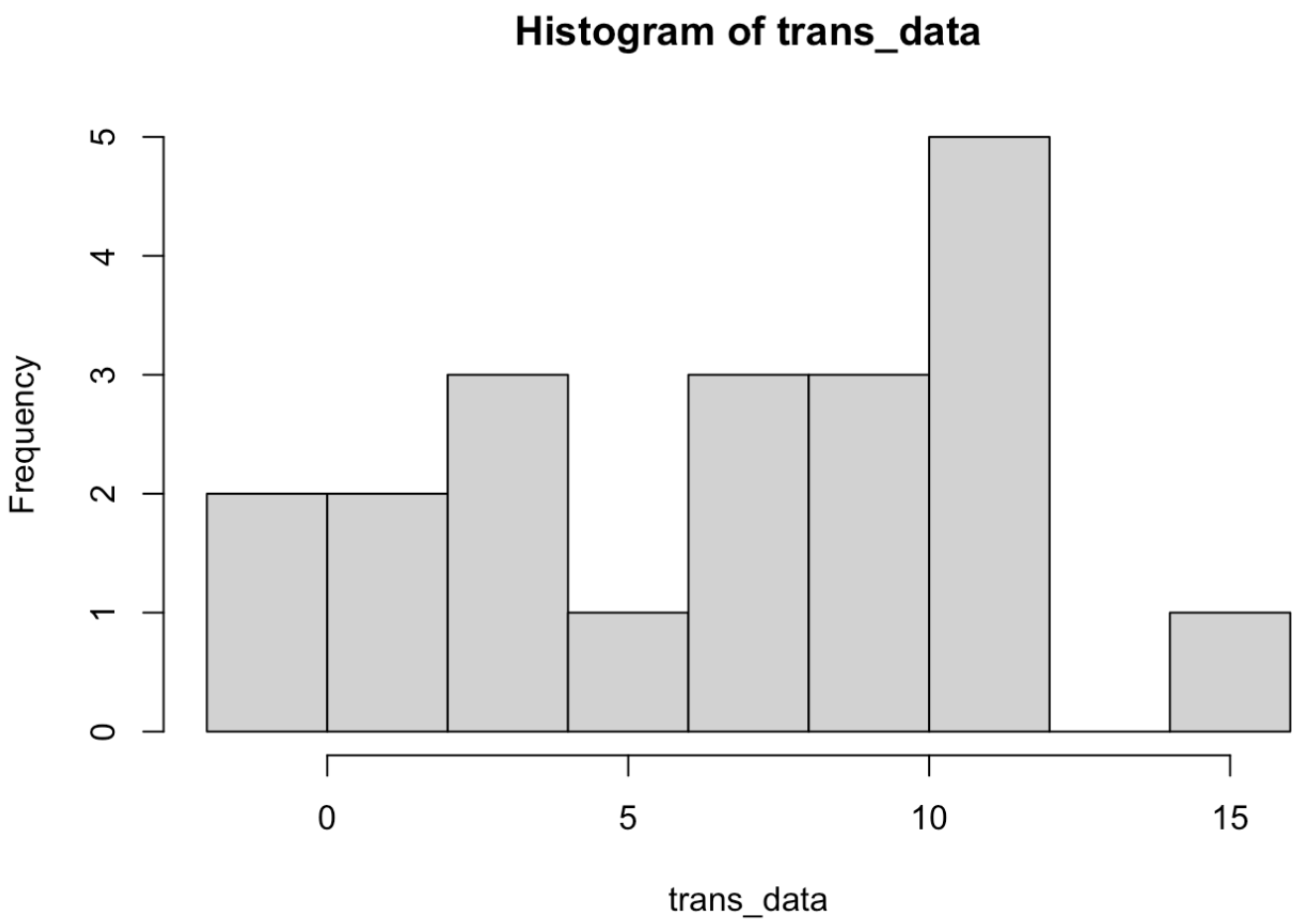


Based on the plot above, the 0 is in the confidence interval of the optimal λ and as the estimation of the parameter is close to 0 in this case, so i think we should to apply the logarithmic transformation of the data.

```
lambda <- b_trans$x[which.max(b_trans$y)]
lambda
```

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

```
trans_data <- (weight ^ lambda - 1) / lambda
hist(trans_data)
```



4.27

Here is the data from textbook

```
Jan<-c(0.15,0.25,0.10,0.20,1.85,1.97,0.80,0.20,0.10,0.50,0.82,0.40,1.80,0.20,1.12,1.8
3,
       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(0.30,0.22,0.10,0.12,0.20,0.10,0.10,0.10,0.10,0.10,0.10,0.17,0.20,2.80,0.85,0.1
0,
       0.10,1.23,0.45,0.30,0.20,1.20,0.10,0.15,0.10,0.20,0.10,0.20,0.35,0.62,0.20,1.2
2,
       0.30,0.80,0.15,1.53,0.10,0.20,0.30,0.40,0.23,0.20,0.10,0.10,0.60,0.20,0.50,0.1
5,
      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
```

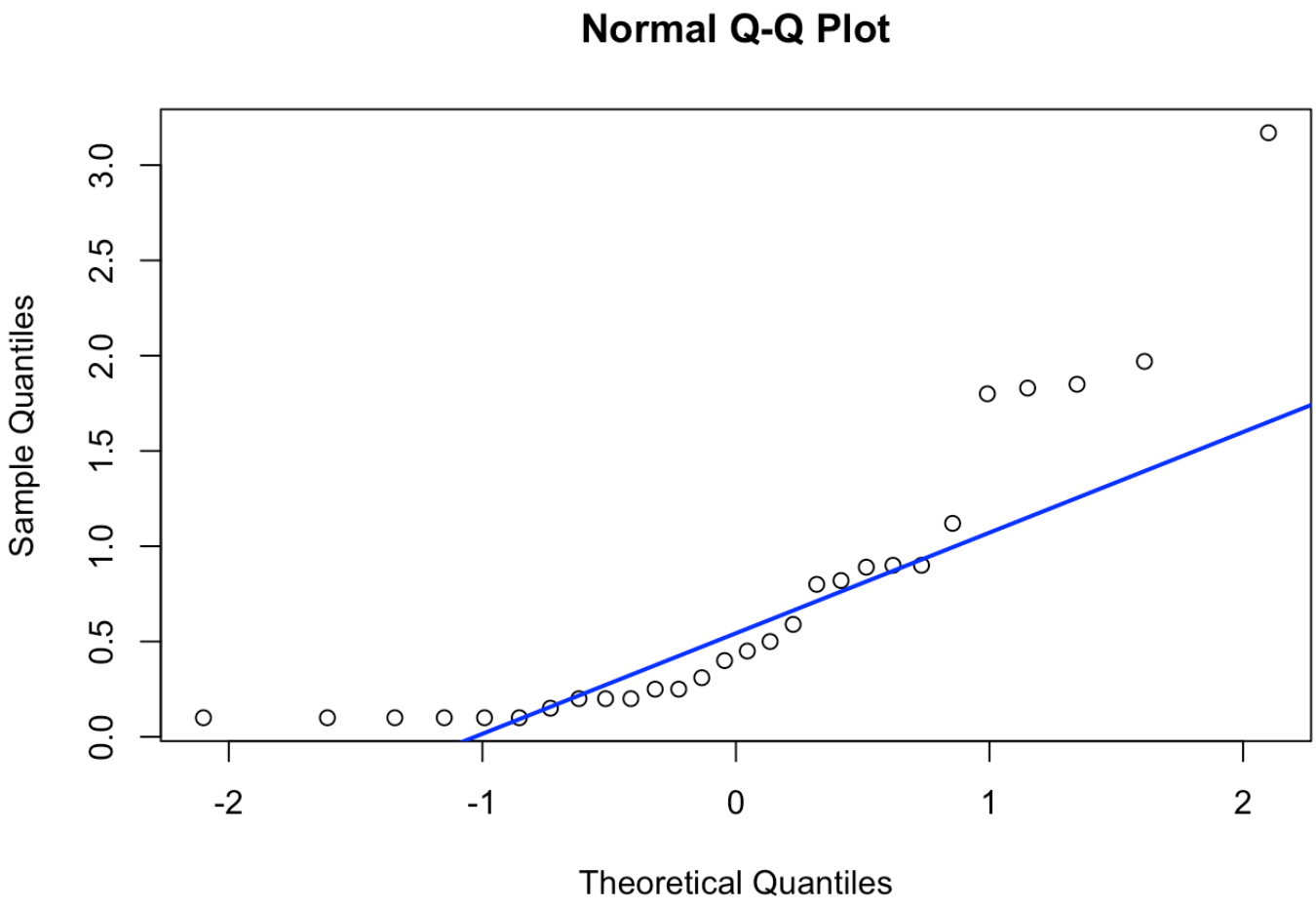
```
summary(Jul)
```

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

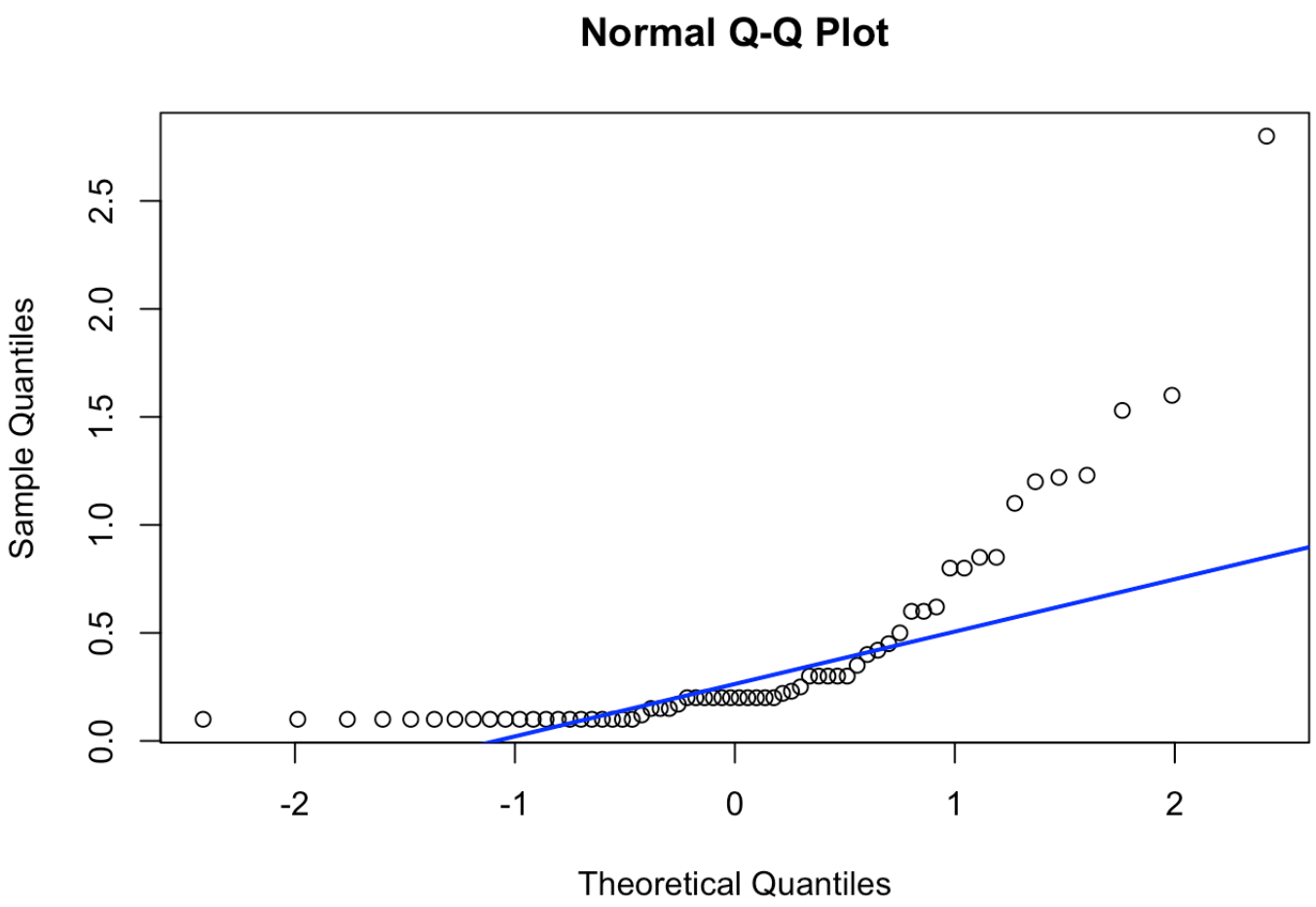
Jan's IQR is higher than the one in Jul, and the mean and median in Jan is larger than Jul.

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

```
#January 1940
qqnorm(Jan, pch = 1)
qqline(Jan, col = "blue", lwd = 2)
```

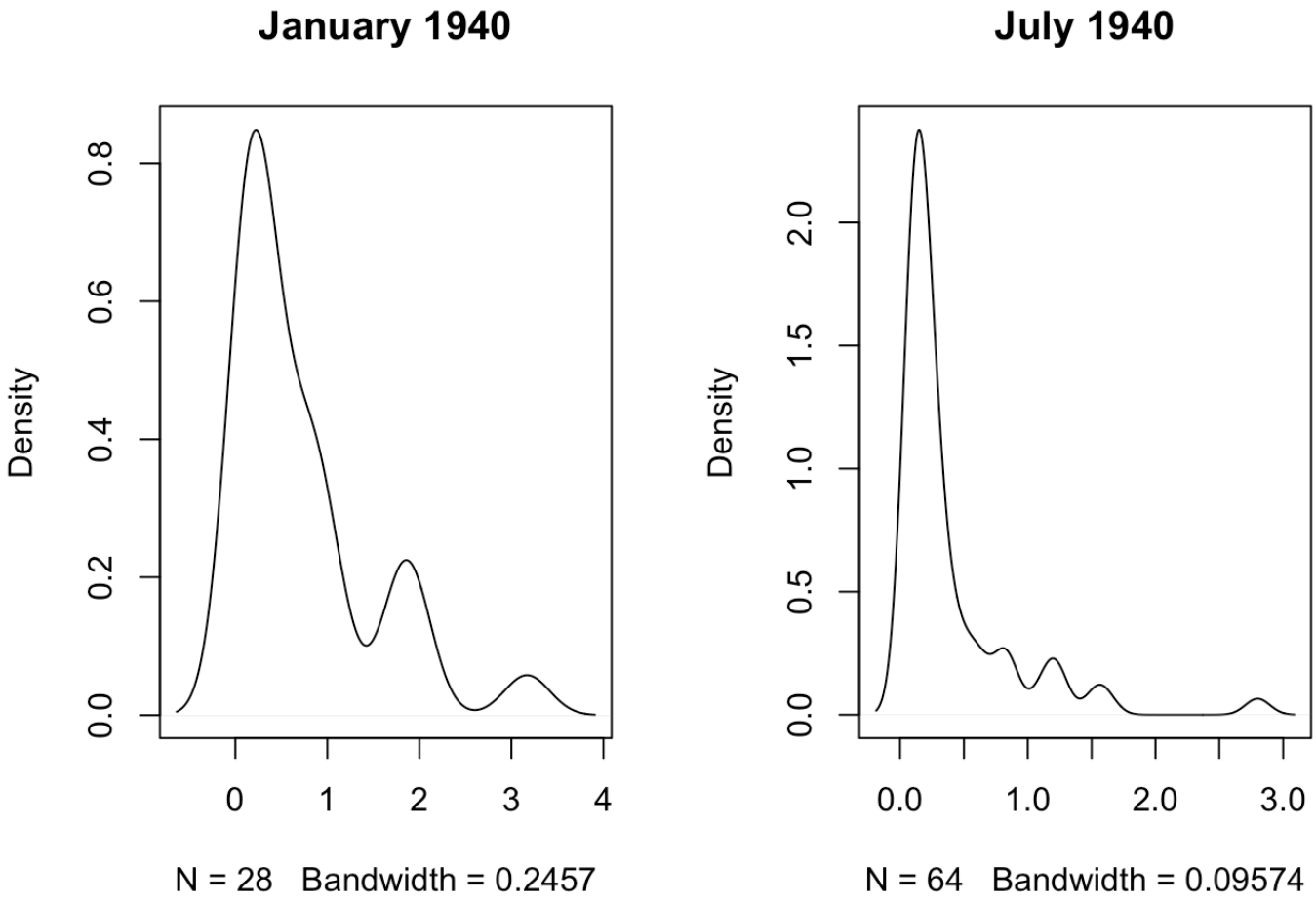


```
#July 1940
qqnorm(Jul, pch = 1)
qqline(Jul, col = "blue", lwd = 2)
```



Density plot

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



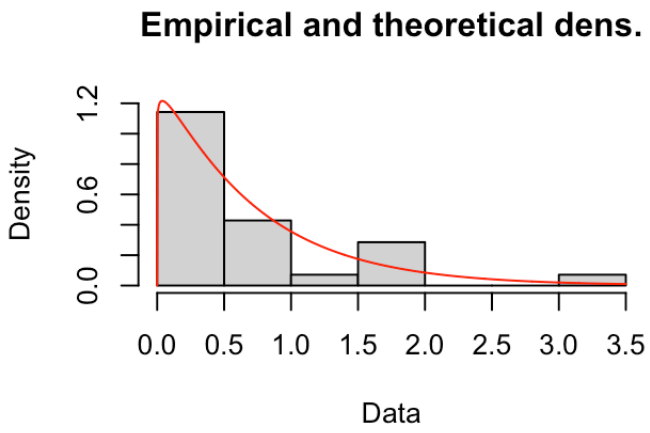
Based on ggplot, the data doesn't follow the normal distribution. Based on density plot, the data looks like gamma distribution, therefore, I suggest gamma distribution.

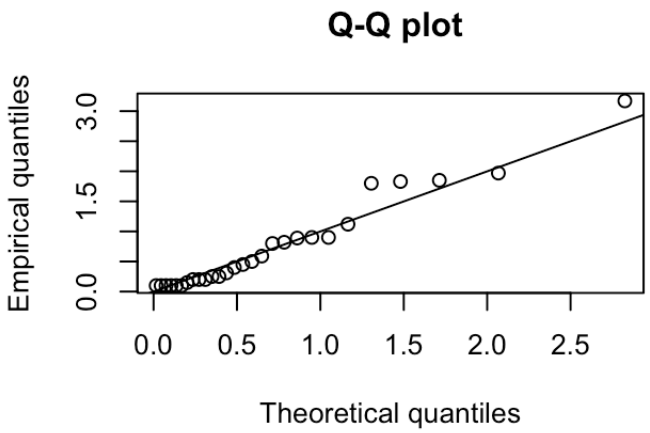
c).

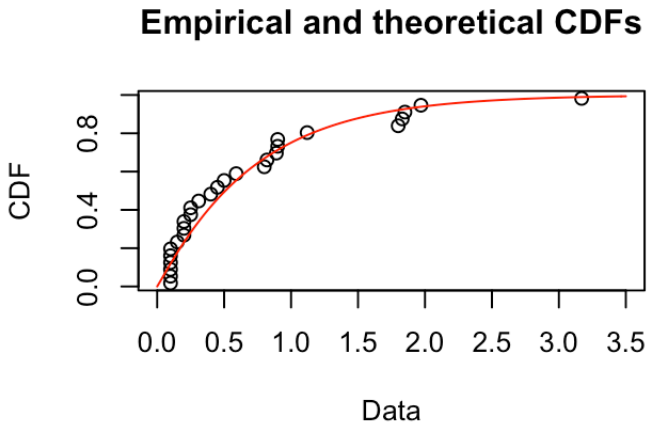
```
Jan.fit <- fitdist(Jan,'gamma','mle')
summary(Jan)
```

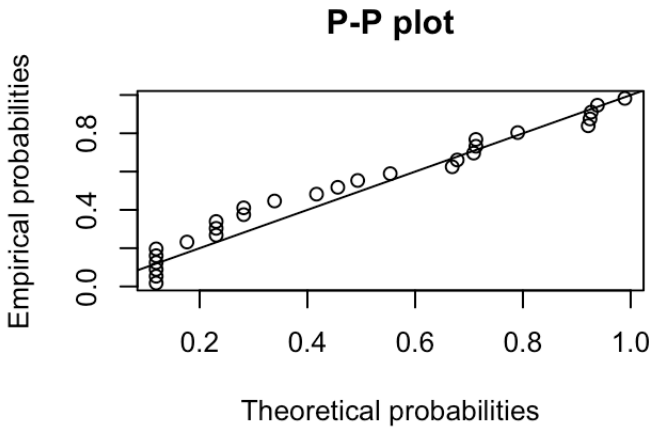
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1875  0.4250  0.7196  0.9000  3.1700
```

```
plot(Jan.fit)
```





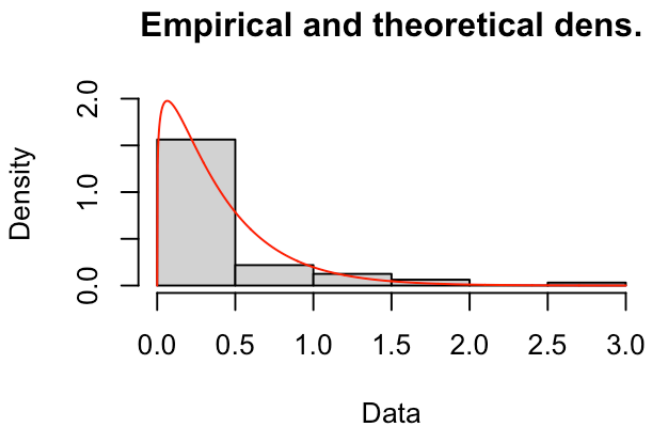


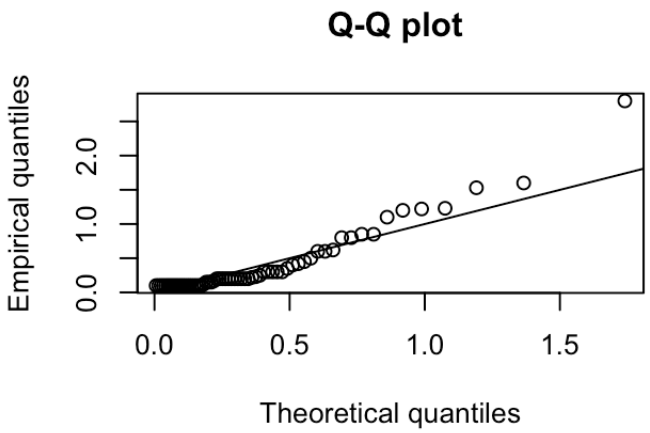


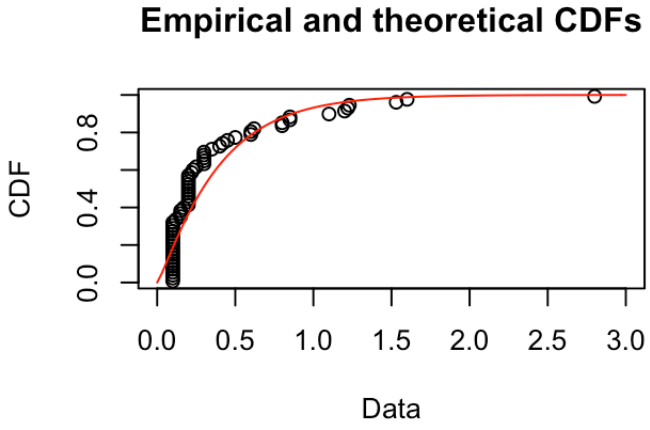

```
July.fit <- fitdist(Jul,'gamma','mle')
summary(July.fit)
```

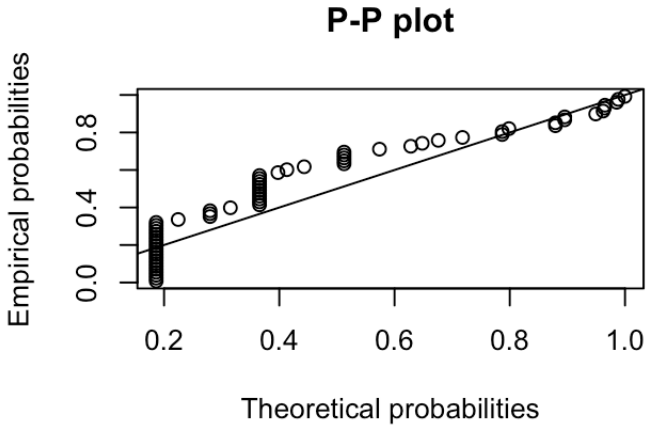
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
## Loglikelihood:  -3.634886   AIC:  11.26977   BIC:  15.58754
## Correlation matrix:
##           shape      rate
## shape 1.0000000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

```
plot(July.fit)
```







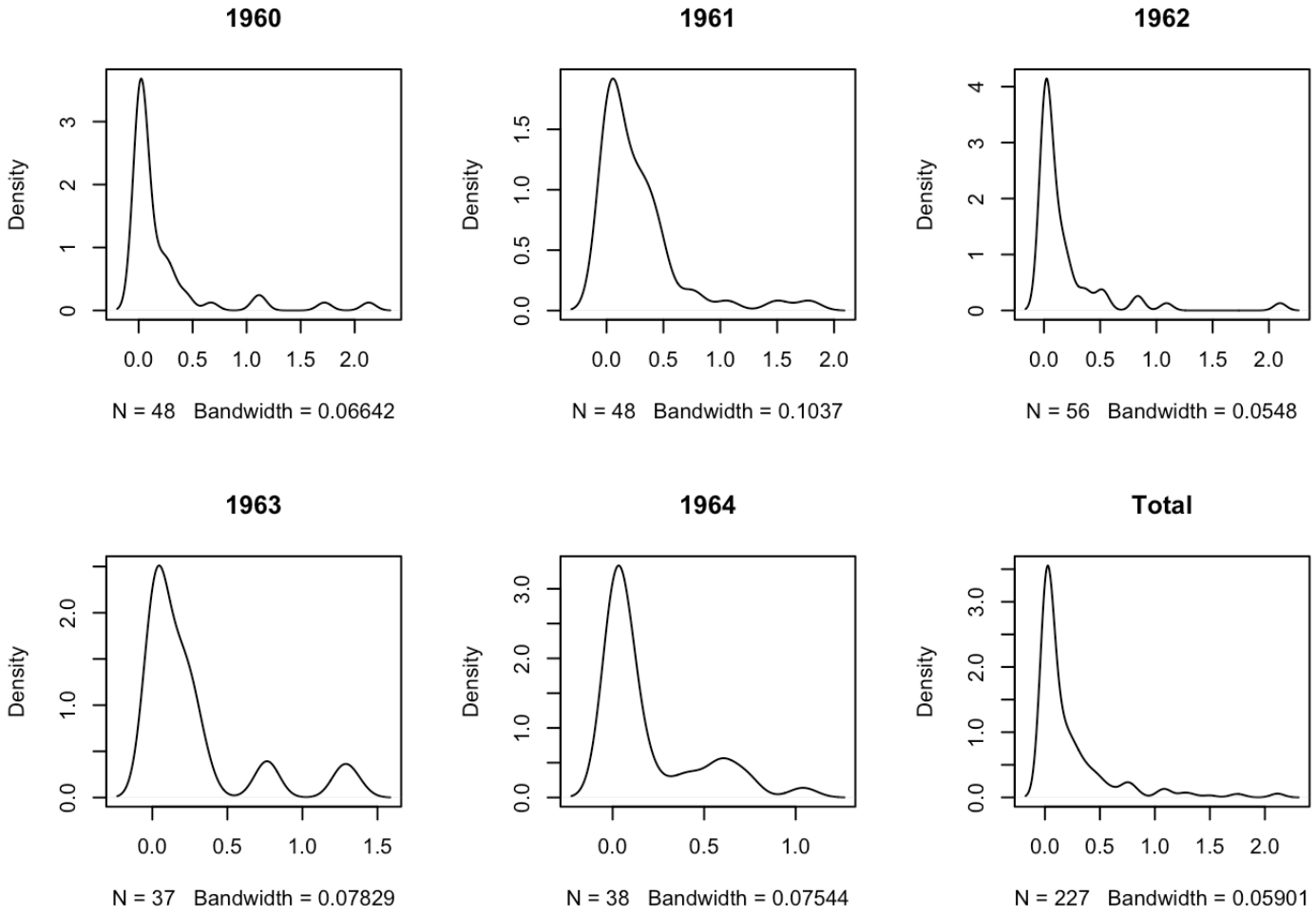


Illinois rain

Q1

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(2).xlsx", sheet = 1, skipEmptyRow
s = FALSE)
par(mfrow = c(2, 3))
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')
```



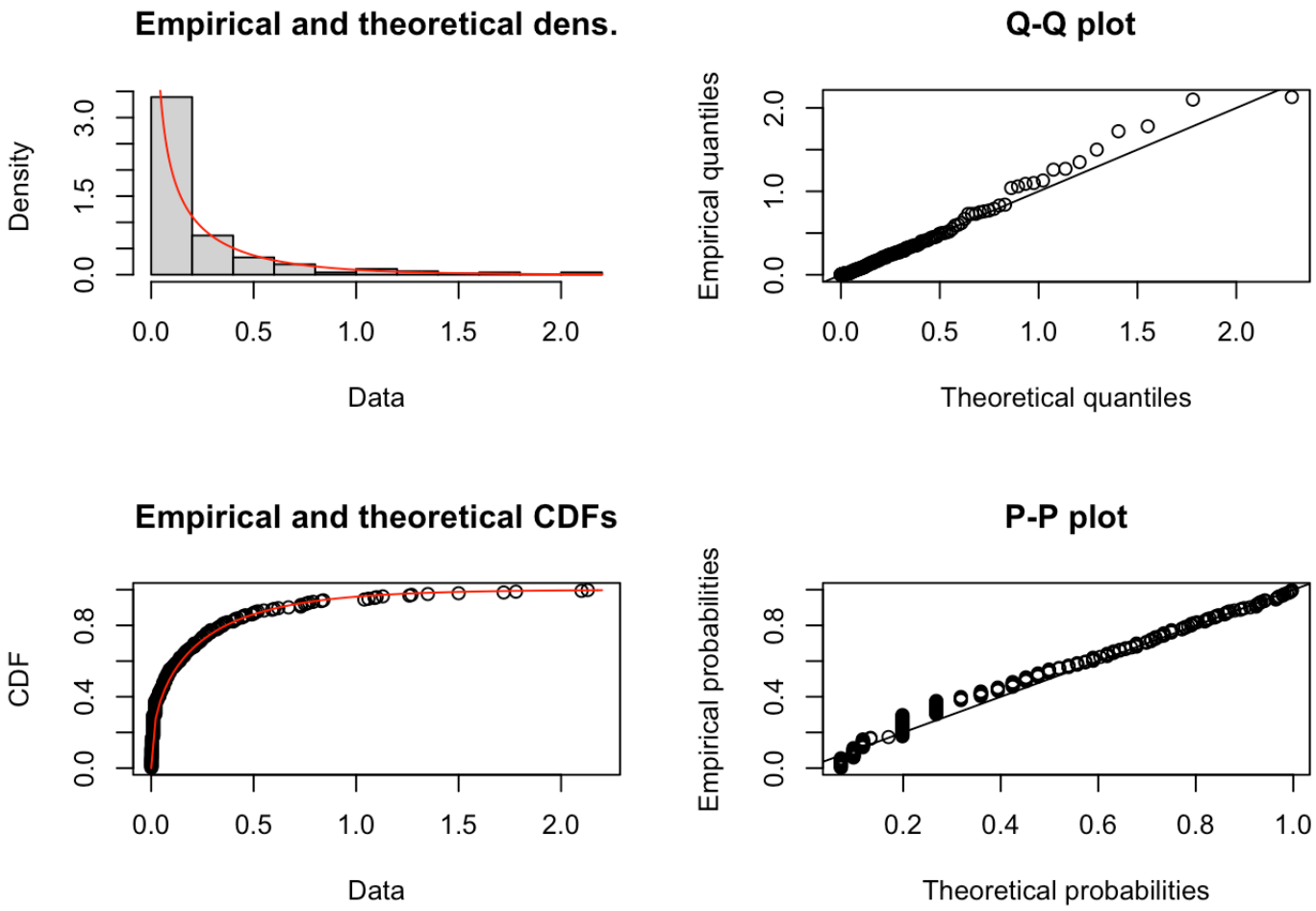
First, I used the whole dataset to conduct fitdist. Gamma distribution is a better choice.

MLE estimation

```
set.seed(2022)
fit1<-fitdist(unlist(rain) %>%  na.omit() %>% c(),'gamma',method='mle')
summary(bootdist(fit1))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4453178 0.3807186 0.5163997
## rate  1.9989008 1.5697775 2.5787418
```

```
plot(fit1)
```



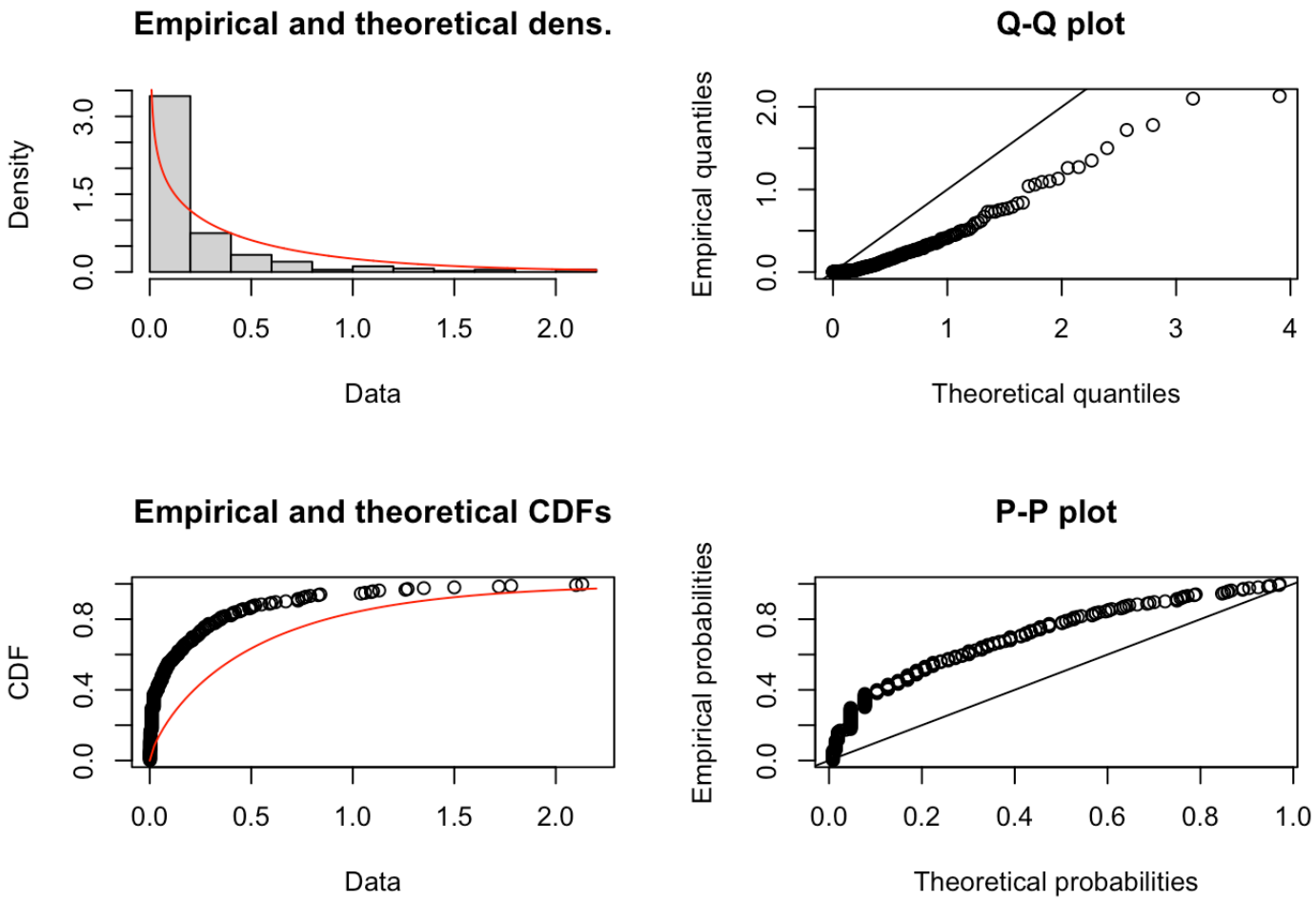
-95% confidence interval: (0.3807186, 0.5163997) -rate: (1.5697775, 2.5787418)

MSE estimation

```
set.seed(2022)
fit2<-fitdist(unlist(rain) %>%  na.omit() %>% c(),'gamma',method='mse')
summary(bootdist(fit2))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.718664 0.6187717 0.8405615
## rate  1.342702 1.0819541 1.6826946
```

```
plot(fit2)
```



-95% confidence interval: (0.6187717, 0.8405615) -rate: (1.0819541, 1.6826946)

The CI indicates that the estimation is reliable. MSE has a narrower CI, so MLE fits the rain data better.

Q2

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?

Average

```
avg <- fit1$estimate[1]/fit1$estimate[2]
yealy_mean <- apply(rain,2,mean,na.rm =TRUE)
storm <- c(yealy_mean,avg %>% as.numeric() %>% round(4))
names(storm)[6]= 'Mean'
#storm
```

Yearly # of storm

```
numofstorm<-c(nrow(rain)-apply(is.na(rain),2,sum))
#numofstorm
#mean(numofstorm) #45.4
```

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

Q3

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

The 5-year data is too small to do the analysis. I think we need to collect more data to validate the result we got. Next step: -Collect the storm rainfall data with more tracking years. -Try to figure out whether gamma distribution is a good fit -Validation

###Reference https://github.com/MA615-Yuli/MA677_final (https://github.com/MA615-Yuli/MA677_final) https://stackoverflow.com/questions/24211595/order-statistics-in-r (https://stackoverflow.com/questions/24211595/order-statistics-in-r) https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.pdf (https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.pdf)