677 final project

Yuchen Liu

2022-05-12

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#4.25
library(orderstats)
os5 <- order_probs(1, 1/16 ,5)
summary(os5)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0003057 0.0003057 0.0003057 0.0003057 0.0003057 0.0003057

os10 <- order_probs(1, 1/31 ,10)
summary(os10)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 1.343e-24 1.343e-24 1.343e-24 1.343e-24 1.343e-24 1.343e-24

#4.27
jan<- c(0.15, 0.25, 0.10, 0.20, 1.85, 1.97, 0.80, 0.20, 0.10, 0.50, 0.8
2, 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, 0.1
0, 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)
```

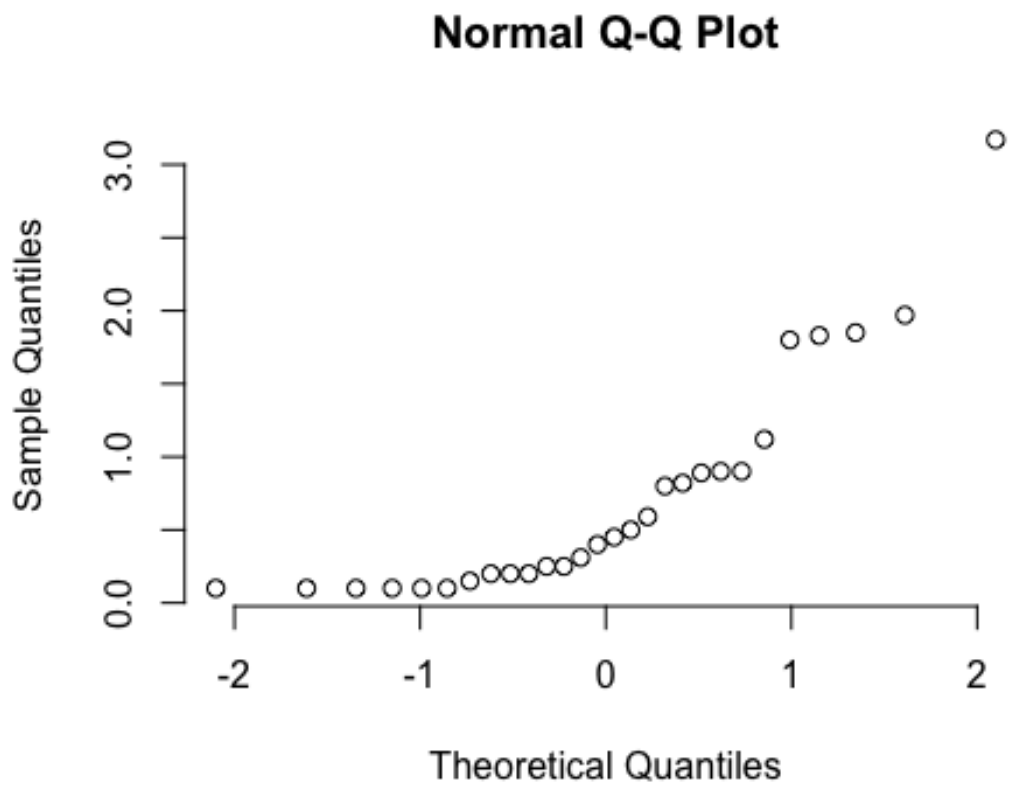
```r
#part a
summary(jan)
```

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

```r
summary(jul)
```

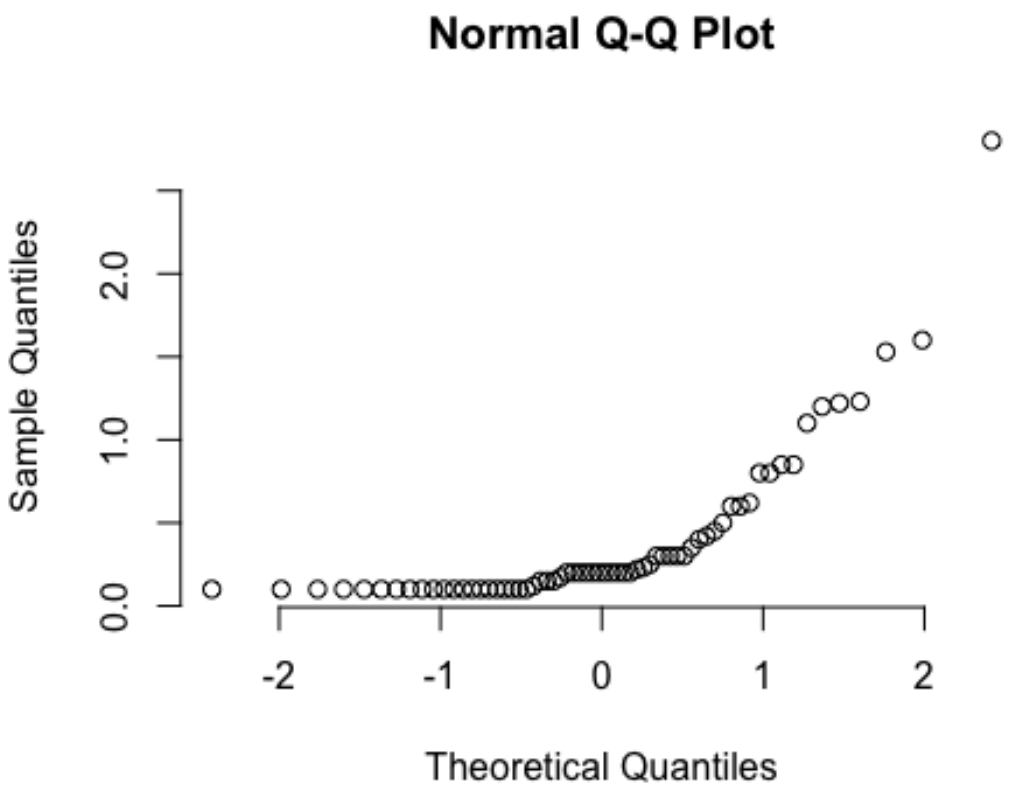
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1000  0.2000  0.3931  0.4275  2.8000
```

```r
# we can see the mean and median in January's rainfall is higher than t
he july. The minimum of january and july are the same but the maximum a
re higher in january.

#part b
janq <- qqnorm(jan, pch = 1, frame = FALSE)
```




```r
julq <- qqnorm(jul, pch = 1, frame = FALSE)

#From the plots, we can see that both of these two datasets do not foll
ow the normal distributions. So we need to use the gamma model.
```

```r
#part c
set.seed(2022)
library(fitdistrplus)

## Loading required package: MASS

## Loading required package: survival
```




```r
library(survival)
library(ProfileLikelihood)
janfg <- fitdist(jan, distr = "gamma", method = "mle")
summary(janfg)

## 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
```

```r
julfg <- fitdist(jul, distr = "gamma", method = "mle")
summary(julfg)
```

```
## 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
```

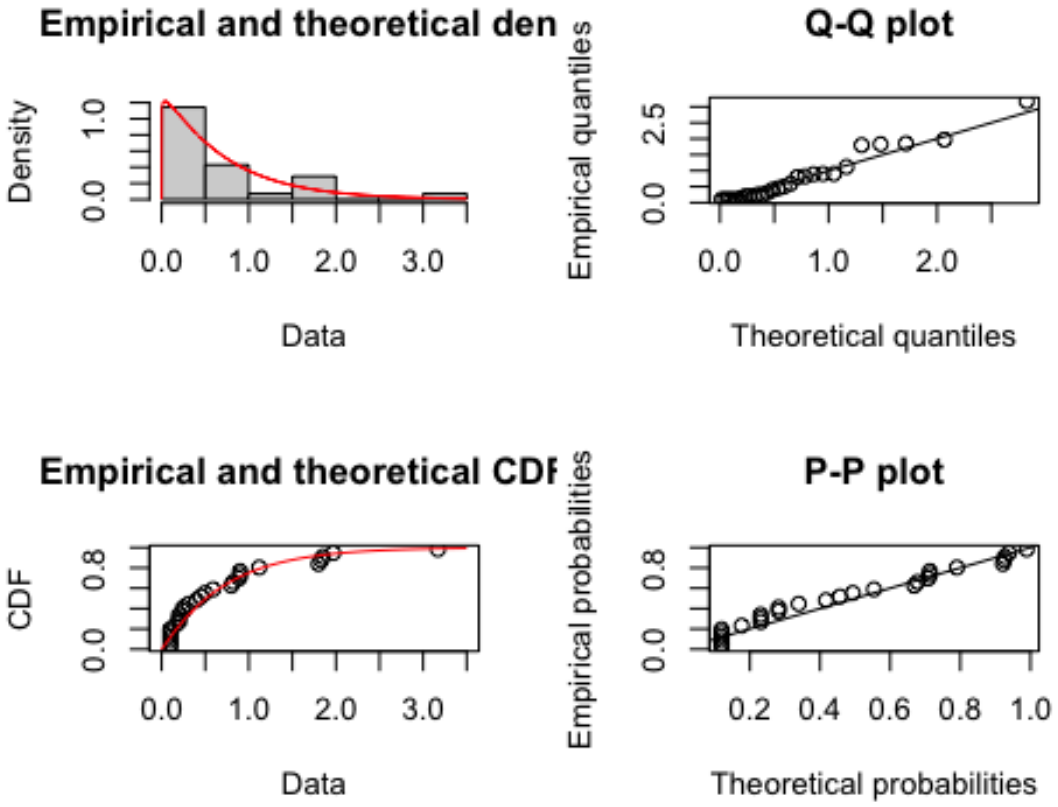
```r
janfg
```

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

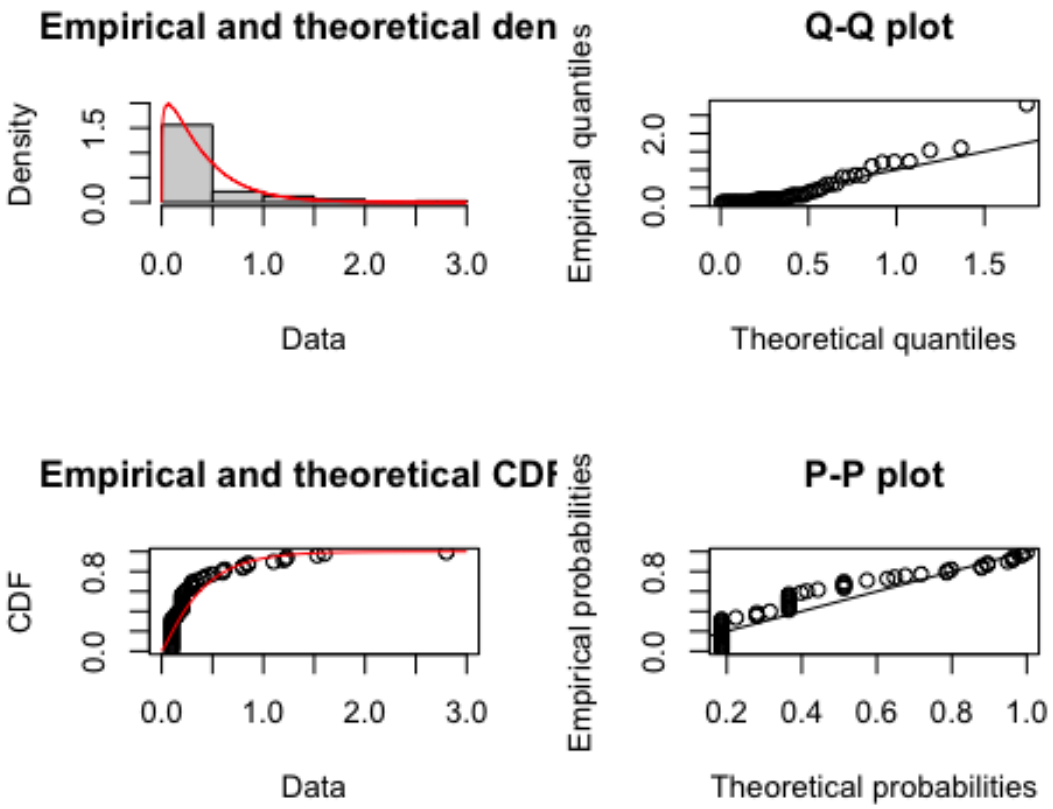
```r
# we can see the results of the MLEs of January dataset is 1.47 and the
 standard error is 0.44.
julfg
```

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

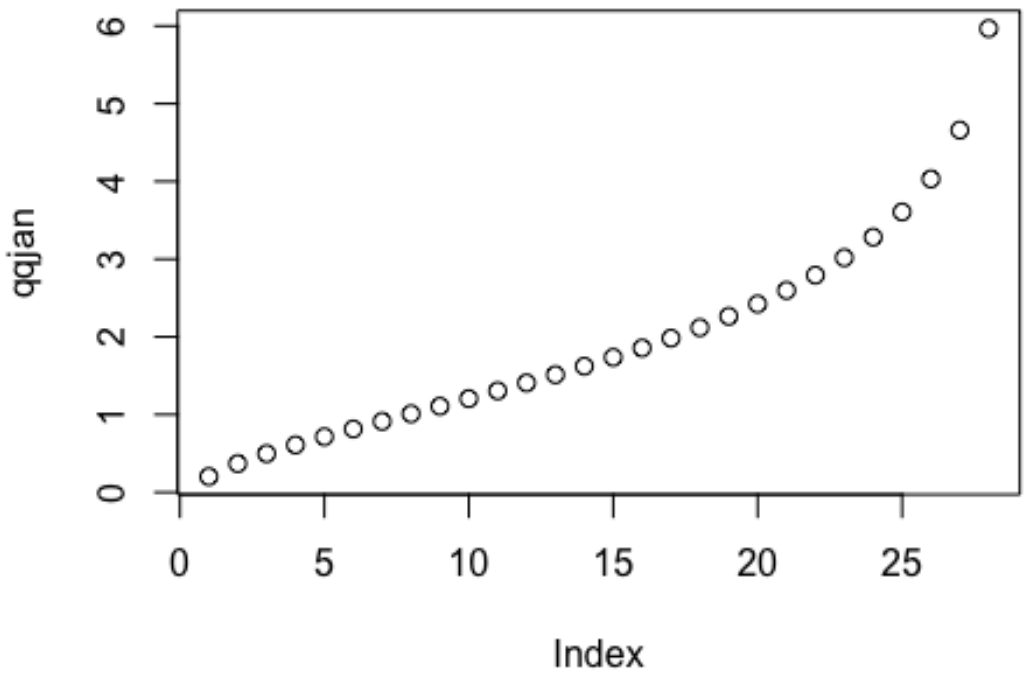
```r
# we can see the results of the MLEs of July dataset is 3.04 and the st
andard error is 0.59.

plot(janfg)
```


Empirical and theoretical den
Density
0.0
1.0
0.0
1.0
2.0
3.0
Data
Q-Q plot
Empirical quantiles
0.0
2.5
0.0
1.0
2.0
Theoretical quantiles
Empirical and theoretical CDF
CDF
0.0
0.8
0.0
1.0
2.0
3.0
Data
P-P plot
Empirical probabilities
0.0
0.8
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities

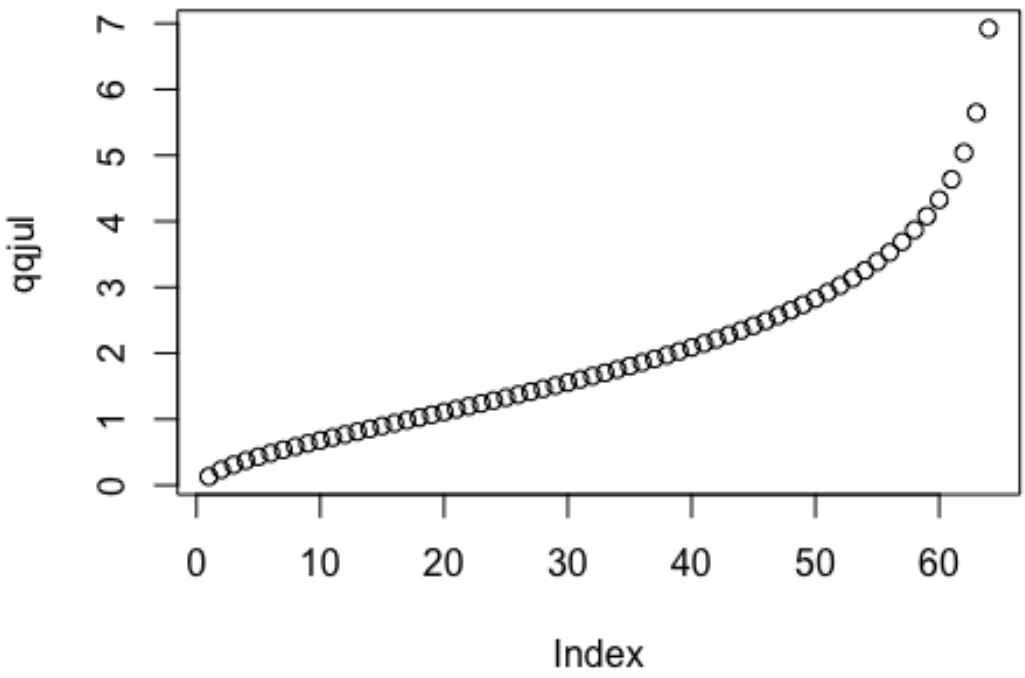
```
plot(julfg)
```



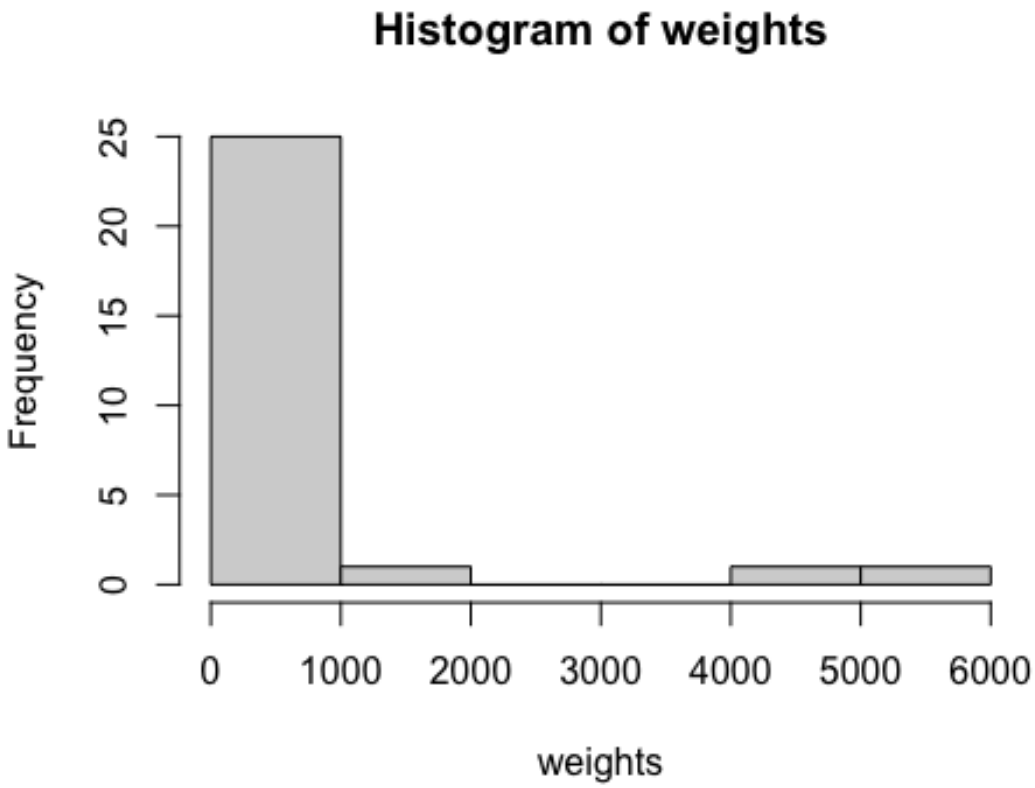

```r
#part d
qqjan <- qgamma(ppoints(length(jan)), shape = 2, rate = 1)
plot(qqjan)
```



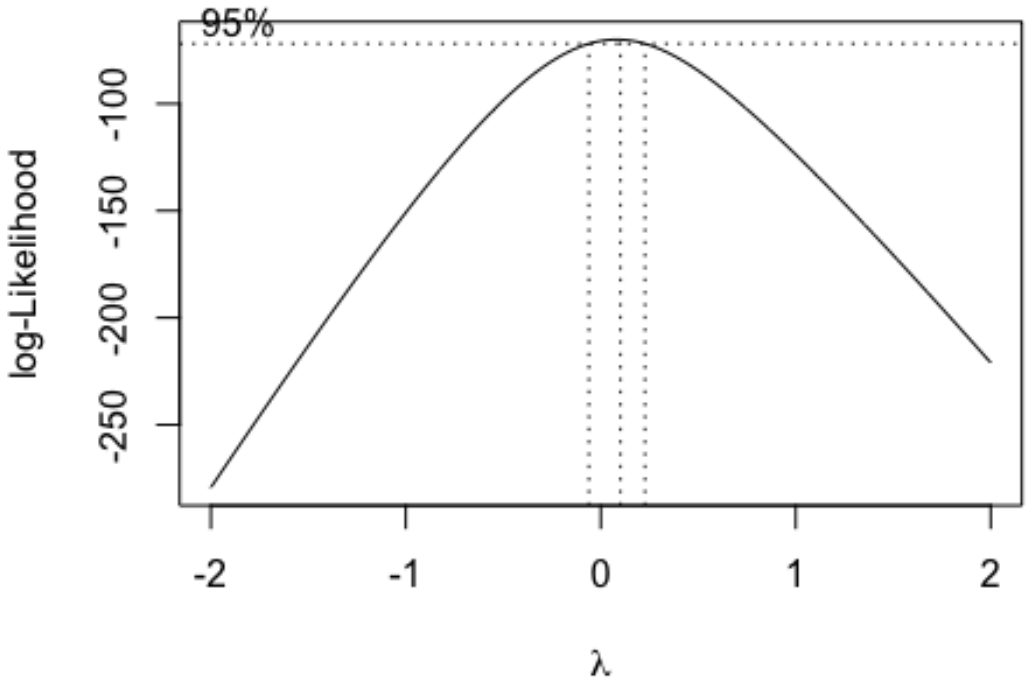

```
qqjul <- qgamma(ppoints(length(jul)), shape = 2, rate = 1)
plot(qqjul)
```



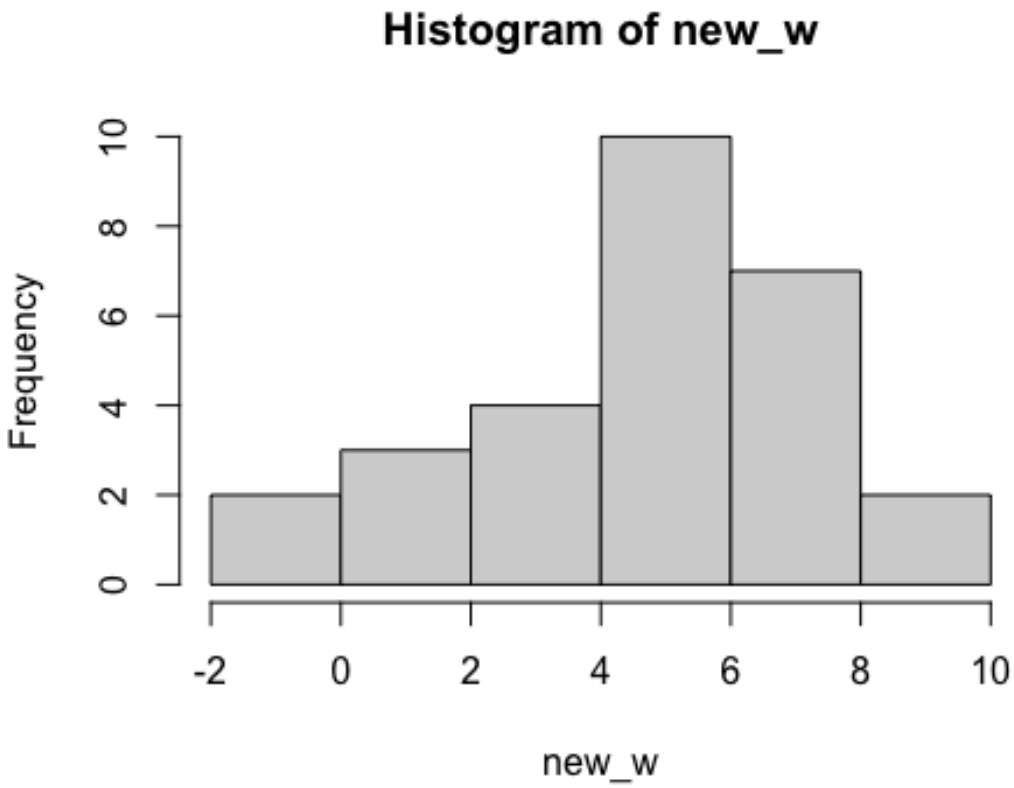

```r
#4.39
library(MASS)
weights<-c(0.4, 1.0, 1.9, 3.0, 5.5, 8.1, 12.1, 25.6, 115.0, 119.5, 154.
5, 157.0, 175.0, 419.0, 423.0, 440.0, 655.0, 680.0, 50.0, 56.0, 70.0, 1
15.0, 179.0, 180.0, 406.0, 1320.0, 4603.0, 5712.0)
hist(weights)
```




```
box<-boxcox(lm(weights ~ 1))
```




```
new_w <- log(weights)
hist(new_w)
```




```
shapiro.test(new_w)

##
##  Shapiro-Wilk normality test
##
## data:  new_w
## W = 0.95787, p-value = 0.31

#rainfall
library(readxl)
library(magrittr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract

library(stats4)
library(maxLik)

## Loading required package: miscTools

##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maxim
um likelihood estimation in R. Computational Statistics 26(3), 443-458.
 DOI 10.1007/s00180-010-0217-1.
##
## If you have questions, suggestions, or comments regarding the 'maxLi
k' package, please use a forum or 'tracker' at maxLik's R-Forge site:
## https://r-forge.r-project.org/projects/maxlik/

rf<-read_excel("rain.xlsx")
rff<-rf %>% pivot_longer(cols = `1960`:`1964`) %>% na.omit()
#rll <- function(mean, log.sd) {-sum(dnorm(rff, mean, exp(log.sd), log=
TRUE))}
#rll <- function(theta) log(theta) - theta*rff
#gradlik <- function(theta) 1/theta - rff
#hesslik <- function(theta) -100/theta^2
#mle <- maxLik(gradlik, start=1, control=list(printLevel=2))
mle <- fitdistr(rff$value, "normal")
est<-mle$estimate
summary(est)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2244  0.2595  0.2947  0.2947  0.3299  0.3650

summary(rff)

##      name               value
##  Length:227         Min.   :0.0010
##  Class :character   1st Qu.:0.0100
##  Mode  :character   Median :0.0700
##                     Mean   :0.2244
##                     3rd Qu.:0.2700
##                     Max.   :2.1300
```

```
#We can see from the estimated paramaters, the mean is 0.2947 and the m
edian is also 0.2947.And I think the distribution is not very accurate.



wet<-filter(rff, rff$value>=0.3299)
summary(wet)

##      name               value
##  Length:49          Min.   :0.330
##  Class :character   1st Qu.:0.420
##  Mode  :character   Median :0.600
##                     Mean   :0.768
##                     3rd Qu.:1.040
##                     Max.   :2.130

sum(wet$name==1960)

## [1] 7

sum(wet$name==1961)

## [1] 16

sum(wet$name==1962)

## [1] 10

sum(wet$name==1963)

## [1] 7

sum(wet$name==1964)

## [1] 9

# we can see that from the dataset, 1961 is the most wet year. There ar
e 16 times in average in this year has heary rainfall from each storm.
And 1962 and 1964 are also wetter than other years. 1963 and 1960 are t
he dryer years. And I think wet years are wet because there were more s
torms in these year.



#After we done the analysis part, I think we should focus more on the r
easons of heary rainfall and do some research of it.

#What I have done in this project and the future plan.

#My coding skills are not good from the begining of our program. Howeve
r, I think I had a improvement after this whole year learning. Although
 it's not good enough than other classmates, right now I can finish a t
horough project by myself. In this project, I do face a lot of problems
```

. Each line of codes have some bugs in the beginning, but I Googled it and get solved it in the end. For the plan of future work is continue to learn and practice coding skills and try to be a good data analyst.