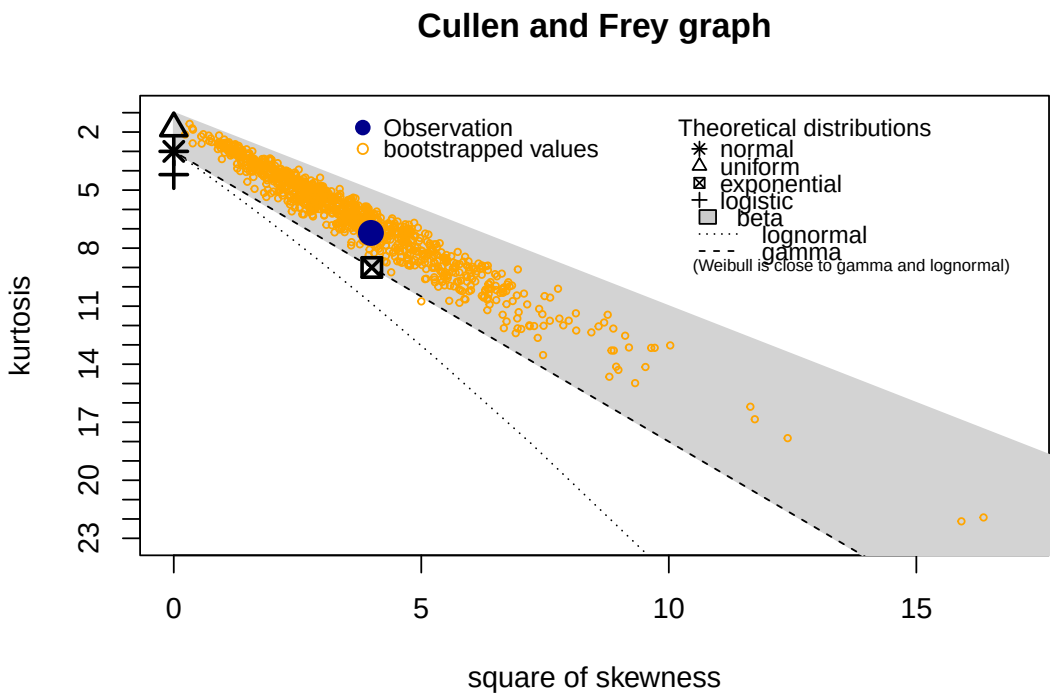
In All Likelihood

Tong Sun

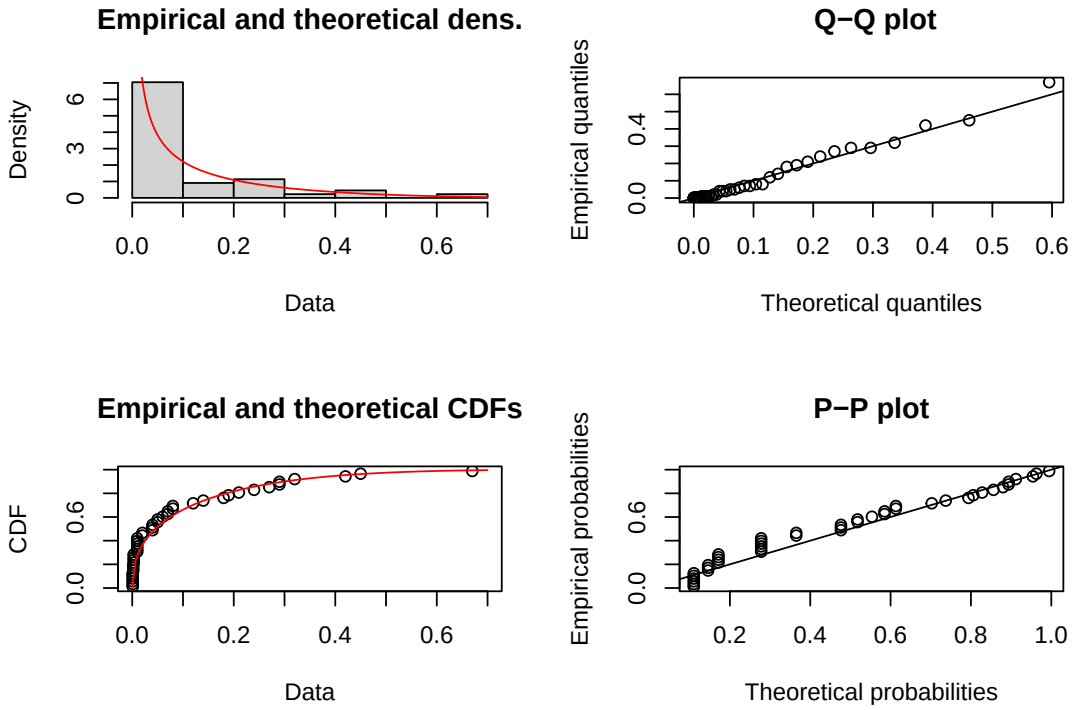
4/24/2022

Rain Data Analysis

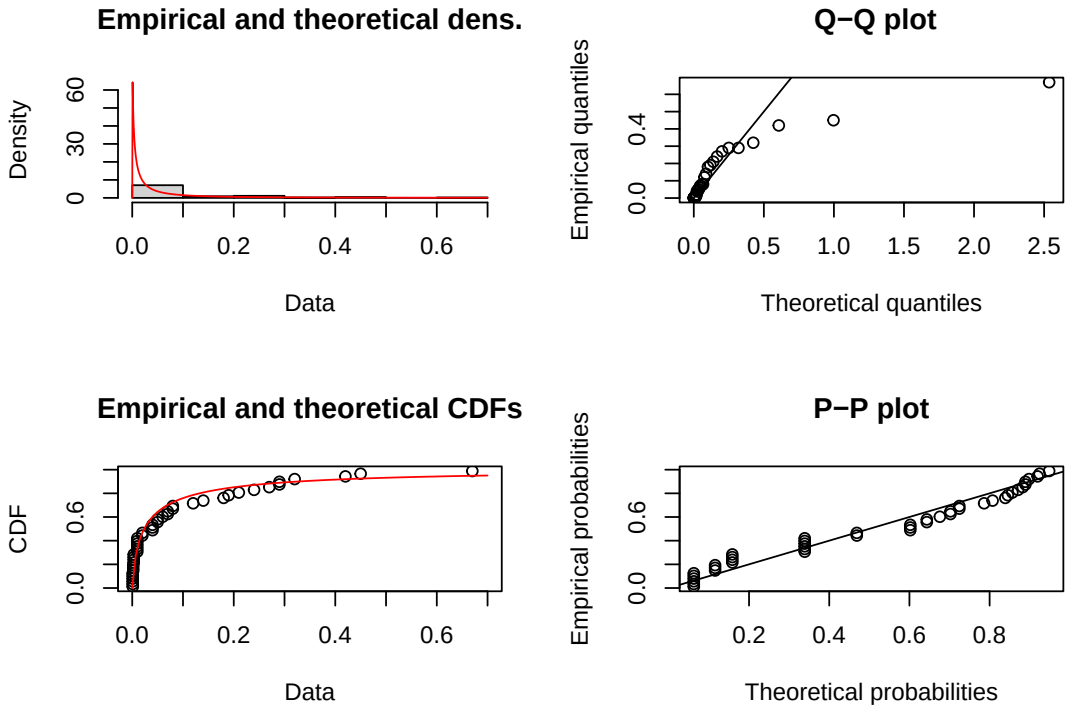
I find an approach of how to identify the distribution of data in R, so I will use it to illustrate this rain problem. A neat approach would involve using "fitdistrplus" package that provides tools for distribution fitting.Details will be shown below.



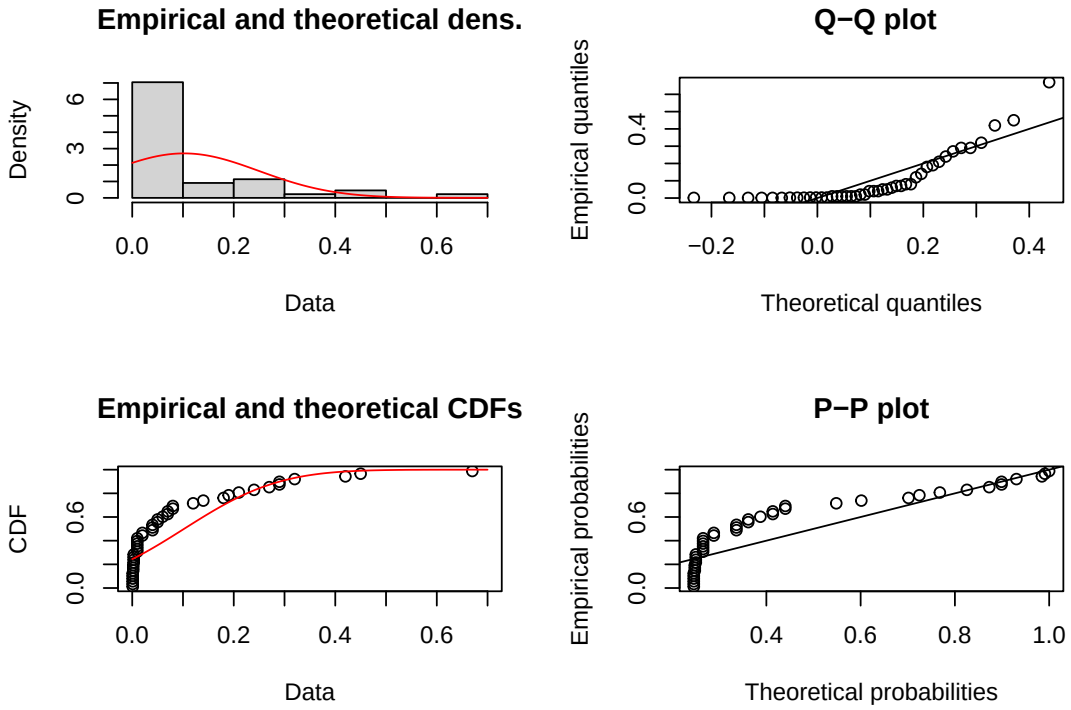

```
## summary statistics
## ------
## min:  0.001   max:  0.67
## median:  0.04
## mean:  0.1021364
## estimated sd:  0.1489462
## estimated skewness:  1.995608
## estimated kurtosis:  7.215314

## Fitting of the distribution ' beta ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape1 0.4044718 0.06997097
## shape2 3.4777921 0.92415999
## Loglikelihood:  70.61875   AIC:  -137.2375   BIC:  -133.6691
## Correlation matrix:
##           shape1    shape2
## shape1 1.0000000 0.5668661
## shape2 0.5668661 1.0000000
```



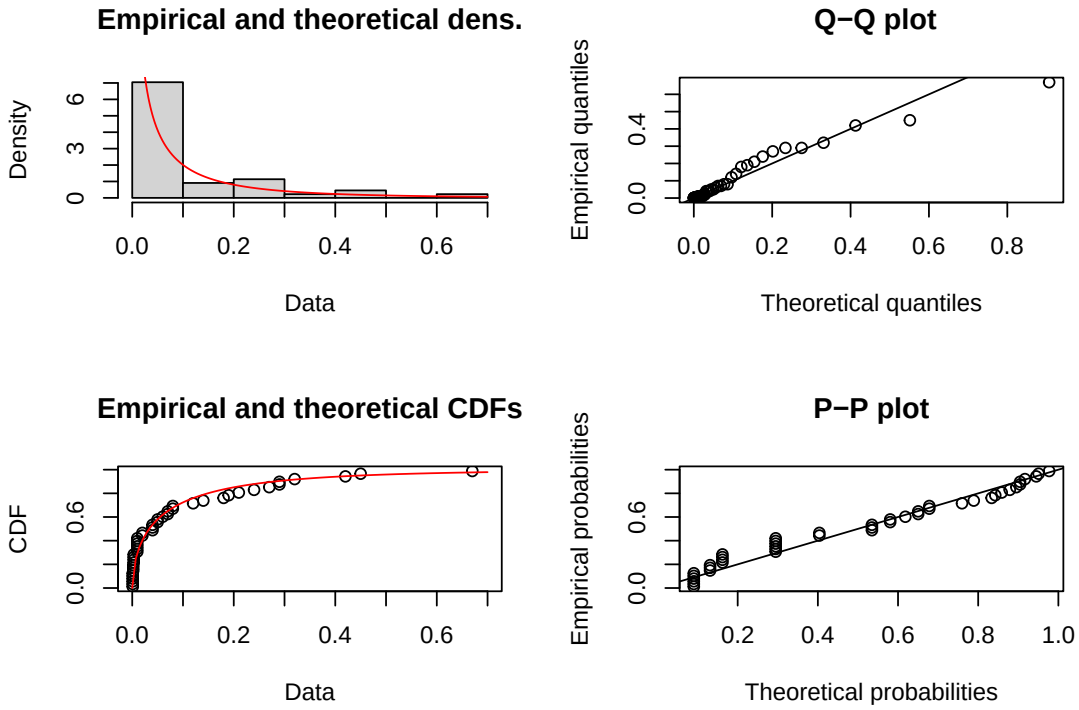

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog -3.751231  0.3098401
## sdlog    2.055247  0.2190898
## Loglikelihood:  70.92343   AIC:  -137.8469   BIC:  -134.2785
## Correlation matrix:
##               meanlog         sdlog
## meanlog  1.000000e+00 -2.411682e-10
## sdlog   -2.411682e-10  1.000000e+00
```




```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 0.1021364 0.02219785
## sd   0.1472439 0.01569299
## Loglikelihood:  21.85597   AIC:  -39.71194   BIC:  -36.14356
## Correlation matrix:
##              mean           sd
## mean 1.000000e+00 3.093975e-13
## sd   3.093975e-13 1.000000e+00
```

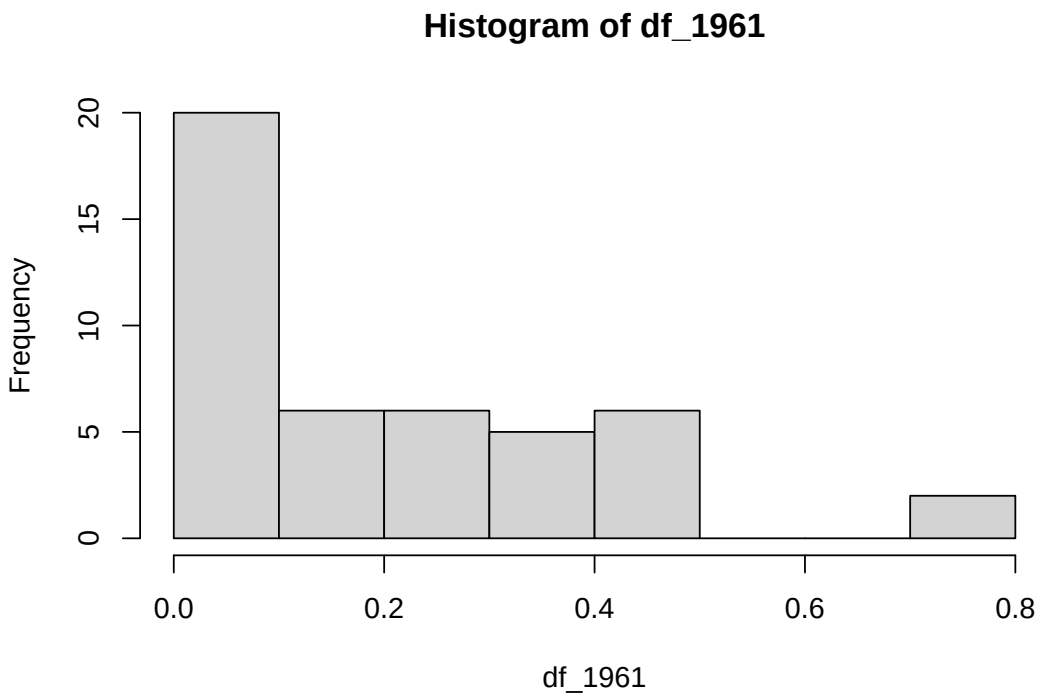


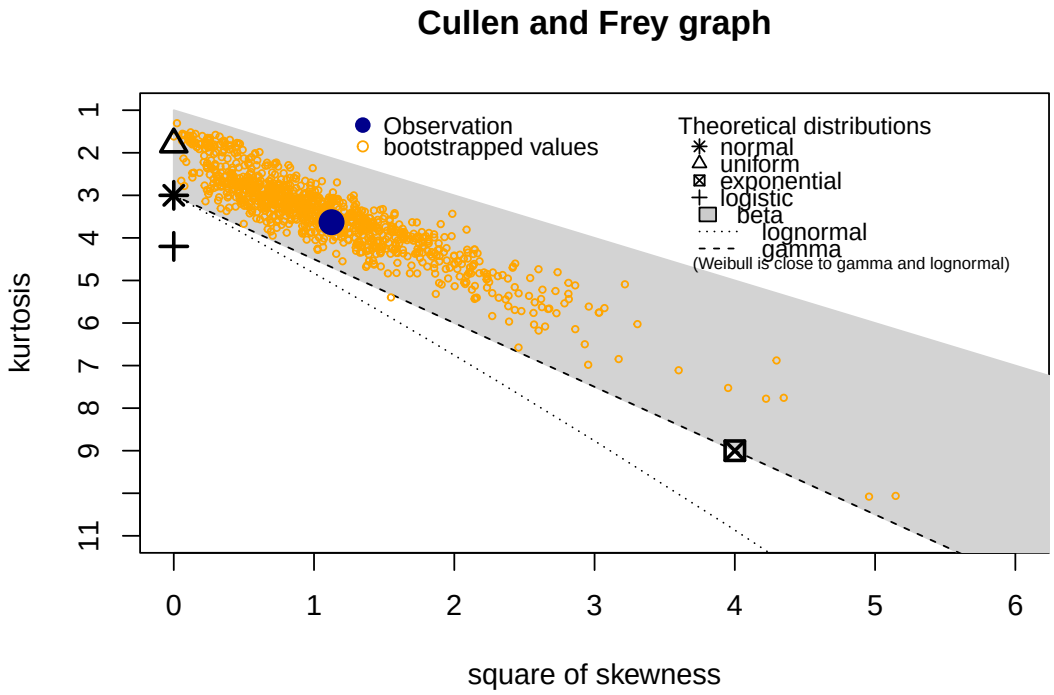

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape 0.56588738 0.06757576
## scale 0.06412626 0.01804239
## Loglikelihood:  70.99206   AIC:  -137.9841   BIC:  -134.4157
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3236002
## scale 0.3236002 1.0000000
```




```
## $distribution
## [1] "Beta"
##
## $sample.size
## [1] 44
##
## $parameters
##    shape1    shape2
## 0.4045193 3.4786494
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mle"
##
## $data.name
## [1] "df_1960"
##
## $bad.obs
## [1] 0
##
## attr(,"class")
## [1] "estimate"
```

From the Cullen and Frey graph, I attempt to fit different distributions. Finally I find the beta distribution fits the data best.Next I use "ebeta" to calculate the parameters with MLE. I got the results that the parameters of this beta distribution is "shape1=0.40" and "shape2=3.48".I also did the same thing on the following years' rain data belowed and I will only show the parameters of each distribution.

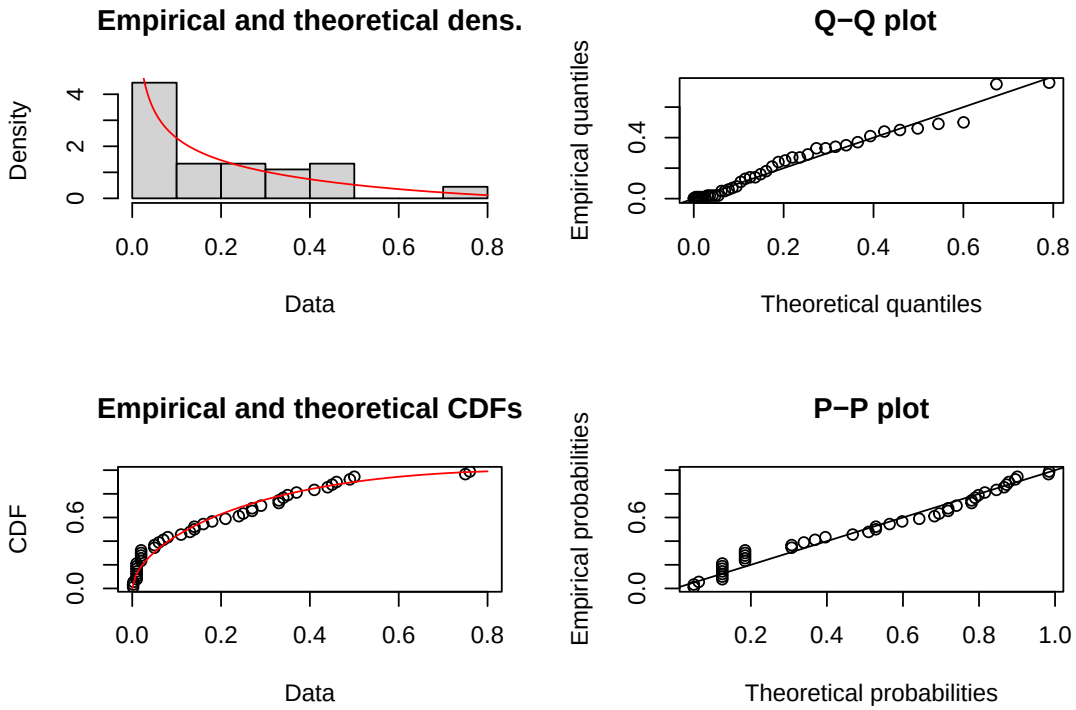




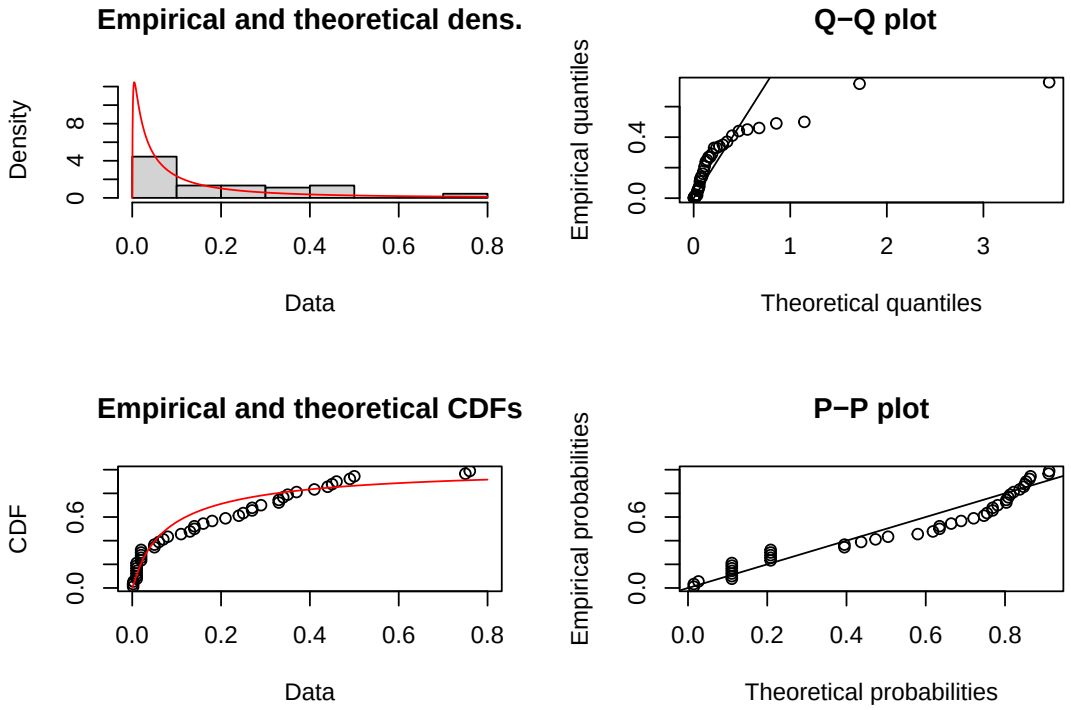

```
## summary statistics
## ------
## min:  0.002   max:  0.76
```

```
## median:  0.14
## mean:  0.1968222
## estimated sd:  0.2021736
## estimated skewness:  1.060626
## estimated kurtosis:  3.630858
```

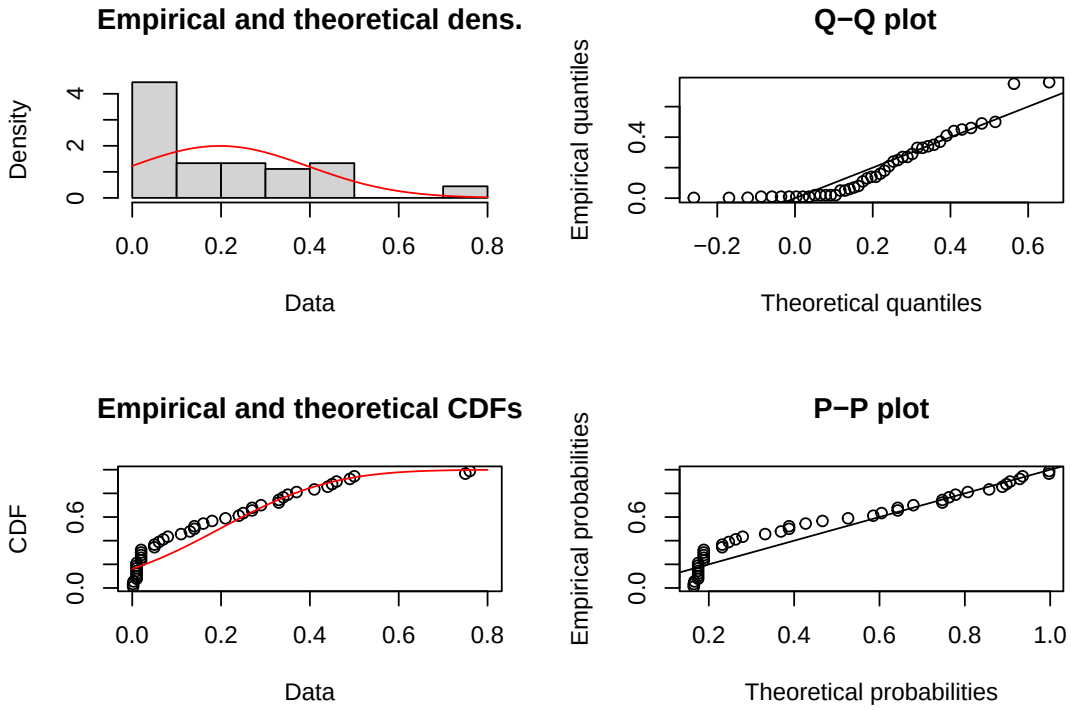
```
## Fitting of the distribution ' beta ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape1 0.5718009  0.1014196
## shape2 2.3683045  0.5522527
## Loglikelihood:  32.85872   AIC:  -61.71744   BIC:  -58.10412
## Correlation matrix:
##           shape1    shape2
## shape1 1.0000000 0.6232764
## shape2 0.6232764 1.0000000
```



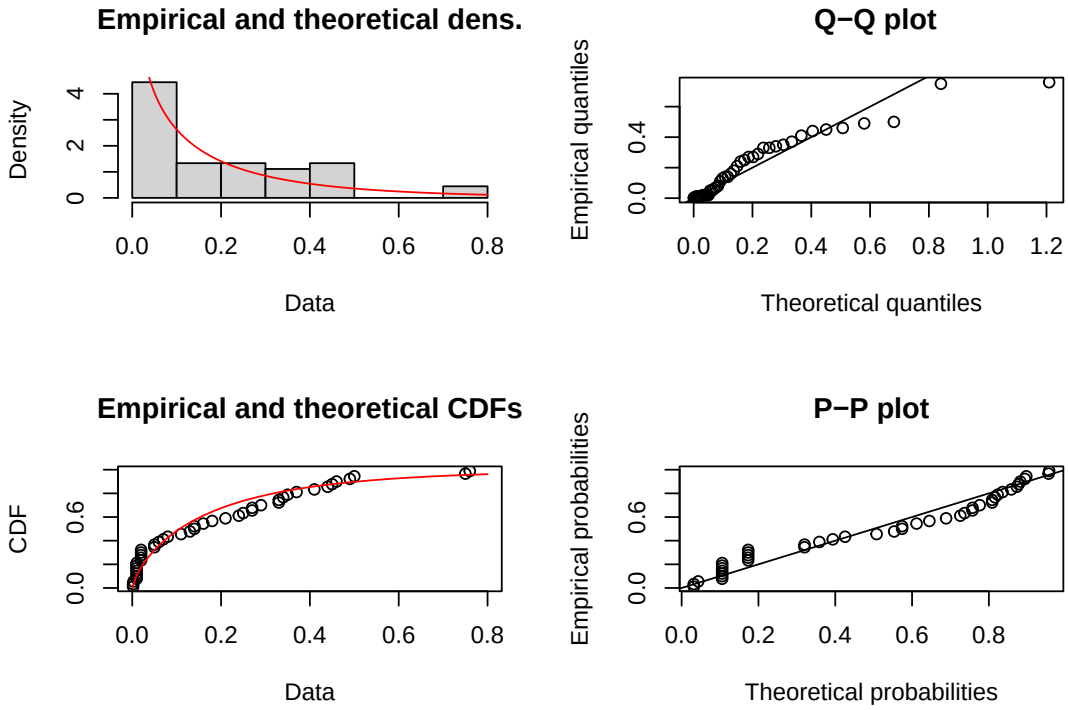

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog -2.546193  0.2509506
## sdlog    1.683428  0.1774486
## Loglikelihood:  27.28899   AIC:  -50.57798   BIC:  -46.96466
## Correlation matrix:
##              meanlog        sdlog
## meanlog 1.000000e+00 5.932699e-11
## sdlog   5.932699e-11 1.000000e+00
```




```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 0.1968222 0.02980150
## sd   0.1999146 0.02107047
## Loglikelihood:  8.591698   AIC:  -13.1834   BIC:  -9.570071
## Correlation matrix:
##               mean            sd
## mean  1.000000e+00 -2.788577e-13
## sd   -2.788577e-13  1.000000e+00
```

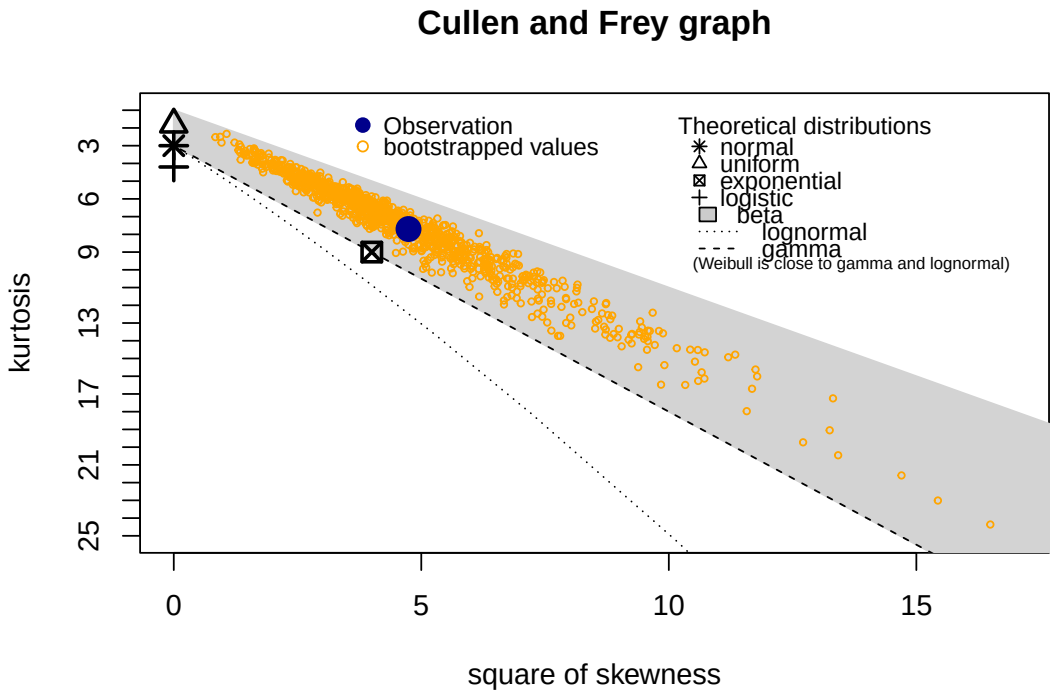



```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.7711396 0.09497226
## scale 0.1719461 0.03495143
## Loglikelihood:  30.6174   AIC:  -57.23479   BIC:  -53.62147
## Correlation matrix:
##          shape    scale
## shape 1.000000 0.309737
## scale 0.309737 1.000000
```




```
## $distribution
## [1] "Beta"
##
## $sample.size
## [1] 45
##
## $parameters
##    shape1    shape2
## 0.5717446 2.3679106
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mle"
##
## $data.name
## [1] "df_1961"
##
## $bad.obs
## [1] 0
##
## attr(,"class")
## [1] "estimate"
```

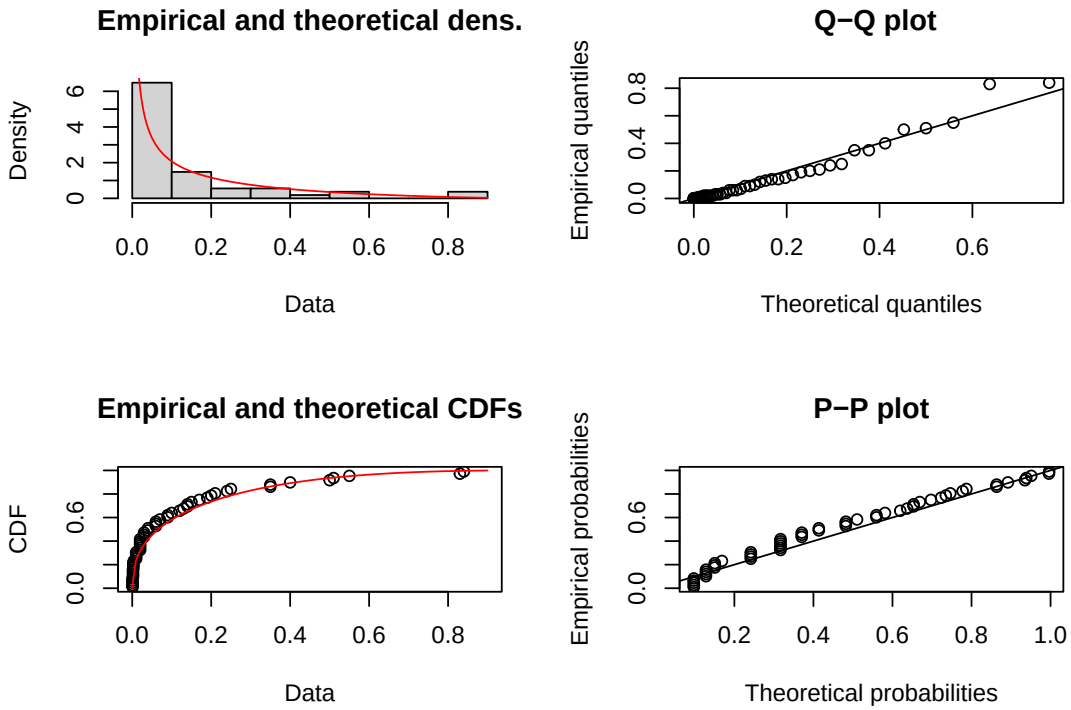
The parameters of 1961 rain data are “shape1=0.57” and “shape2=2.37”.



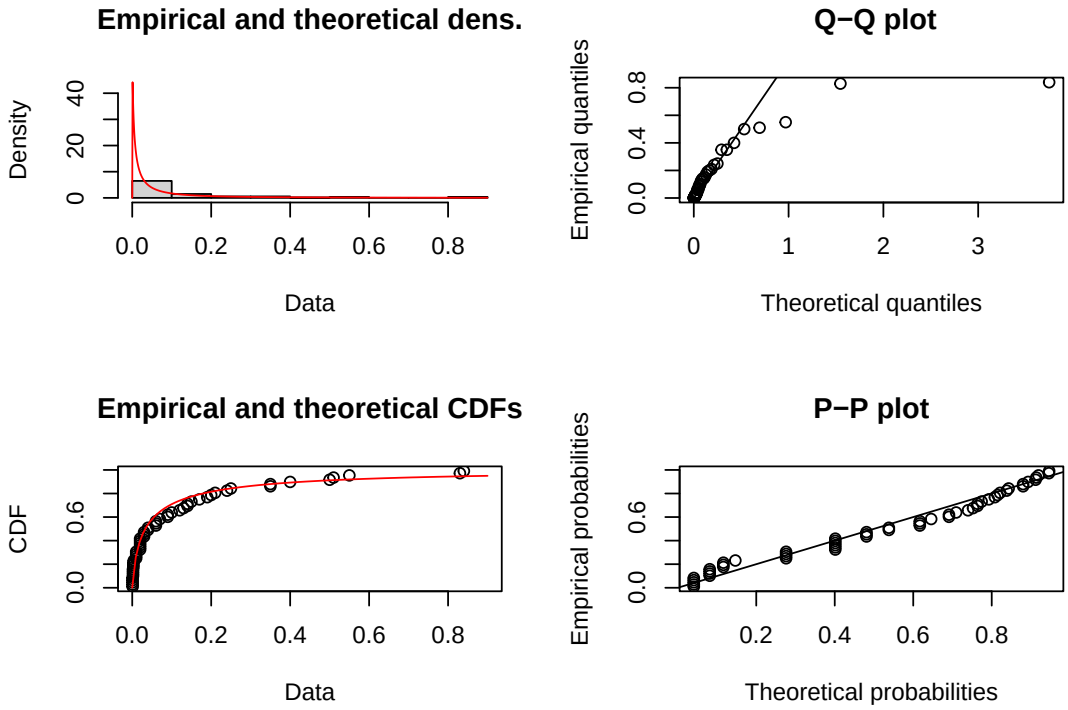

```
## summary statistics
## ------
## min:  0.001   max:  0.84
```

```
## median:  0.04
## mean:  0.1325185
## estimated sd:  0.1979452
## estimated skewness:  2.177623
## estimated kurtosis:  7.699481
```

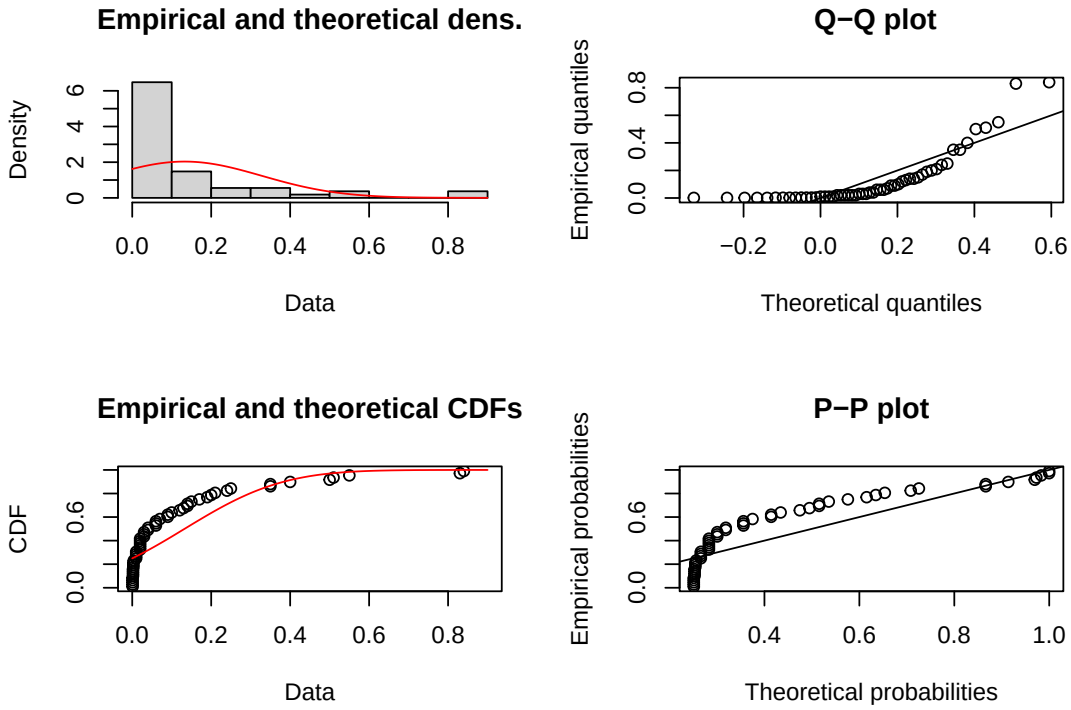
```
## Fitting of the distribution ' beta ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape1 0.3961024 0.06188932
## shape2 2.3575118 0.54582602
## Loglikelihood:  69.32953   AIC:  -134.6591   BIC:  -130.6811
## Correlation matrix:
##           shape1    shape2
## shape1 1.0000000 0.5531286
## shape2 0.5531286 1.0000000
```



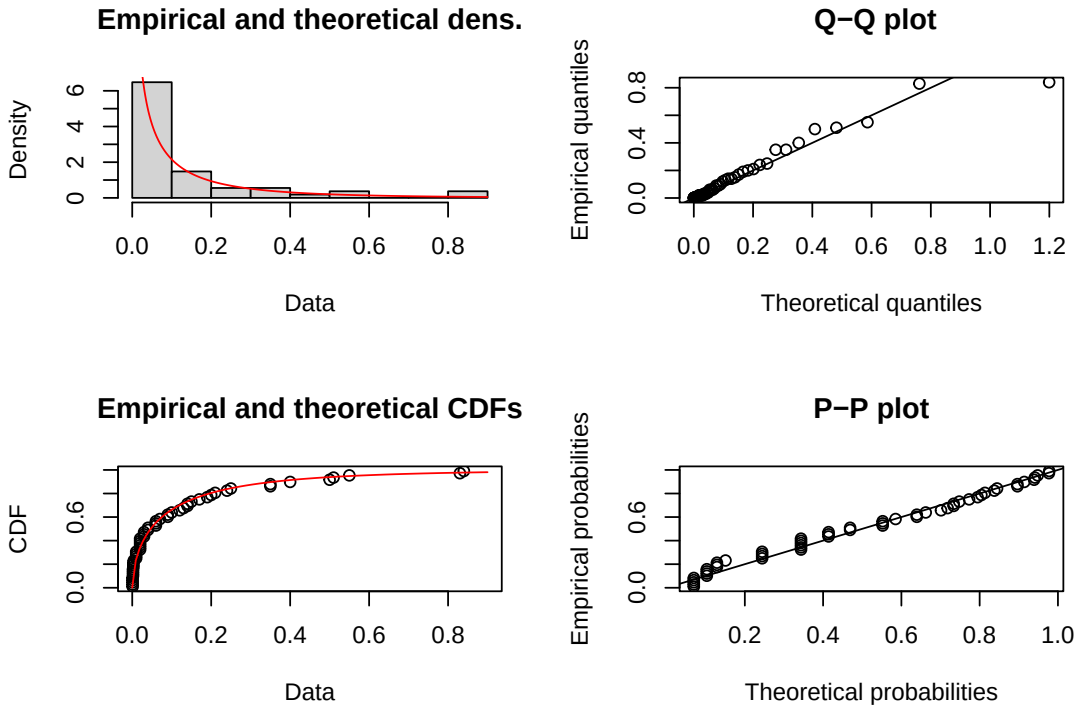

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog -3.410394  0.2734126
## sdlog    2.009164  0.1933317
## Loglikelihood:  69.86178   AIC:  -135.7236   BIC:  -131.7456
## Correlation matrix:
##         meanlog sdlog
## meanlog       1     0
## sdlog         0     1
```




```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 0.1325185 0.02668634
## sd   0.1961038 0.01886789
## Loglikelihood:  11.34933   AIC:  -18.69865   BIC:  -14.72069
## Correlation matrix:
##              mean           sd
## mean 1.000000e+00 4.472111e-13
## sd   4.472111e-13 1.000000e+00
```

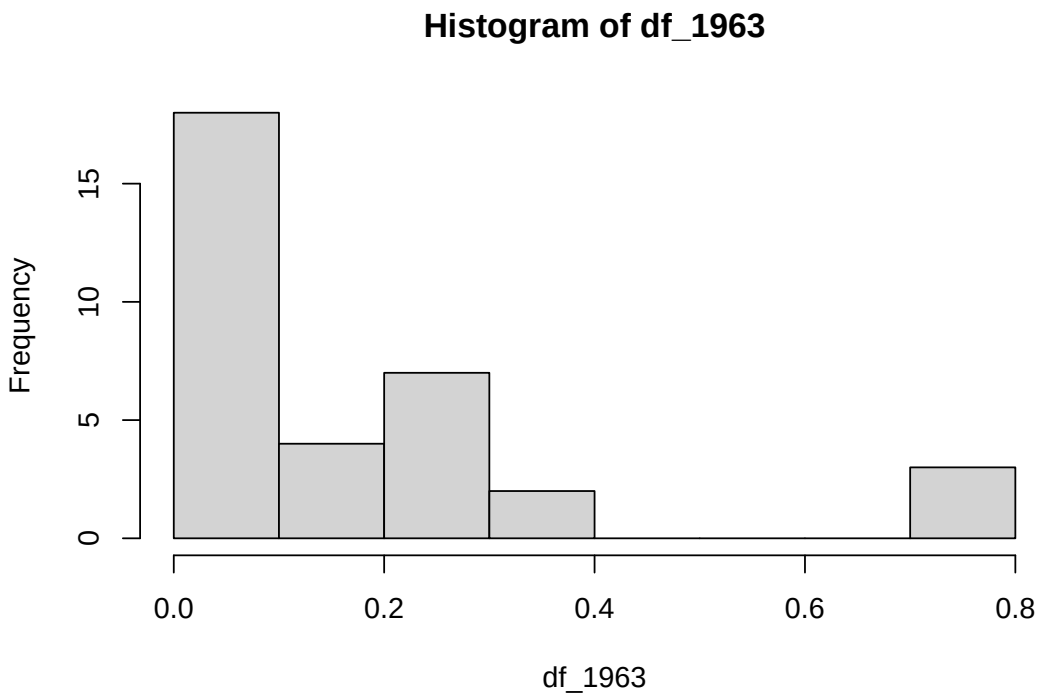


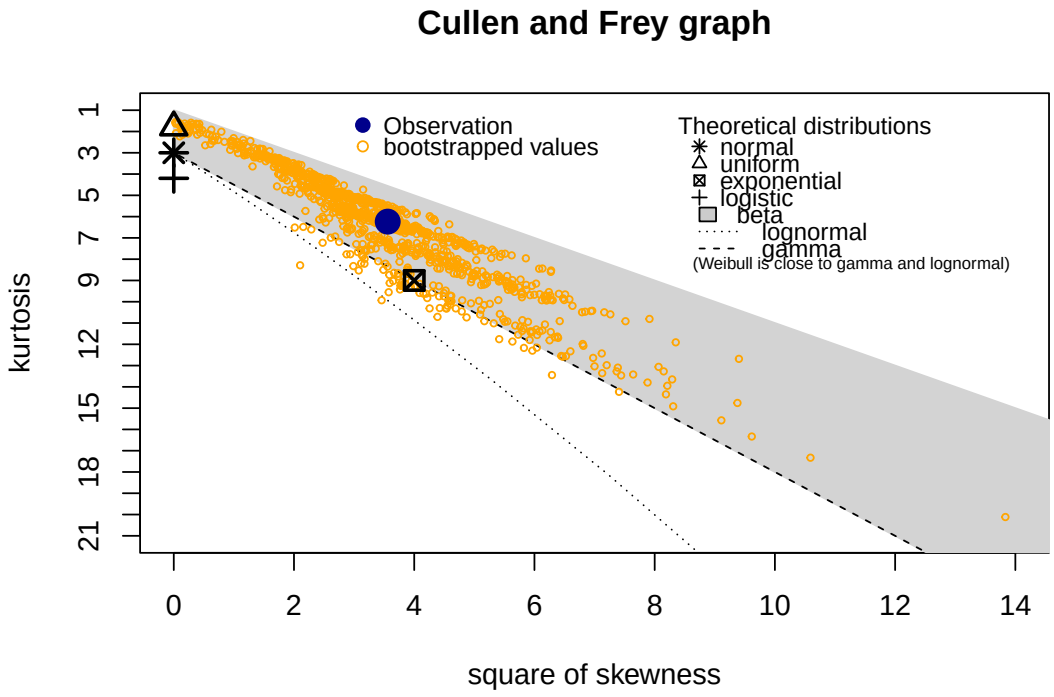

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape 0.58842663 0.06319044
## scale 0.08700623 0.02123960
## Loglikelihood:  70.74253   AIC:  -137.4851   BIC:  -133.5071
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3214015
## scale 0.3214015 1.0000000
```




```
## $distribution
## [1] "Beta"
##
## $sample.size
## [1] 54
##
## $parameters
##    shape1    shape2
## 0.3961229 2.3576651
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mle"
##
## $data.name
## [1] "df_1962"
##
## $bad.obs
## [1] 0
##
## attr(,"class")
## [1] "estimate"
```

The parameters of 1962 rain data are “shape1=0.40” and “shape2=2.36”.

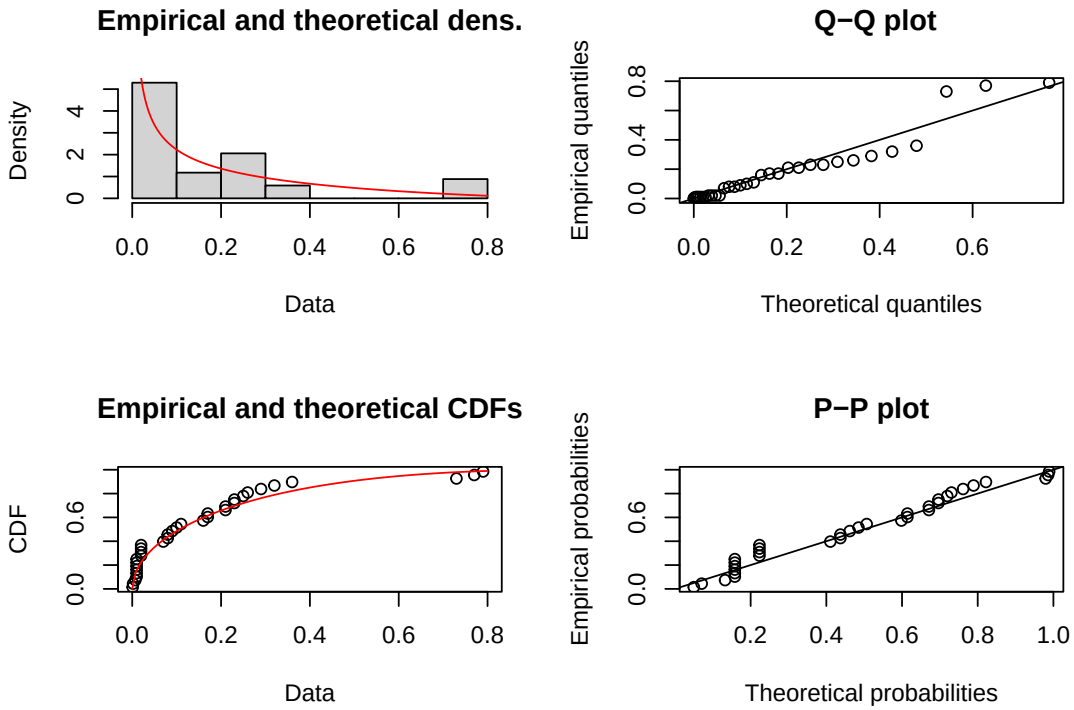




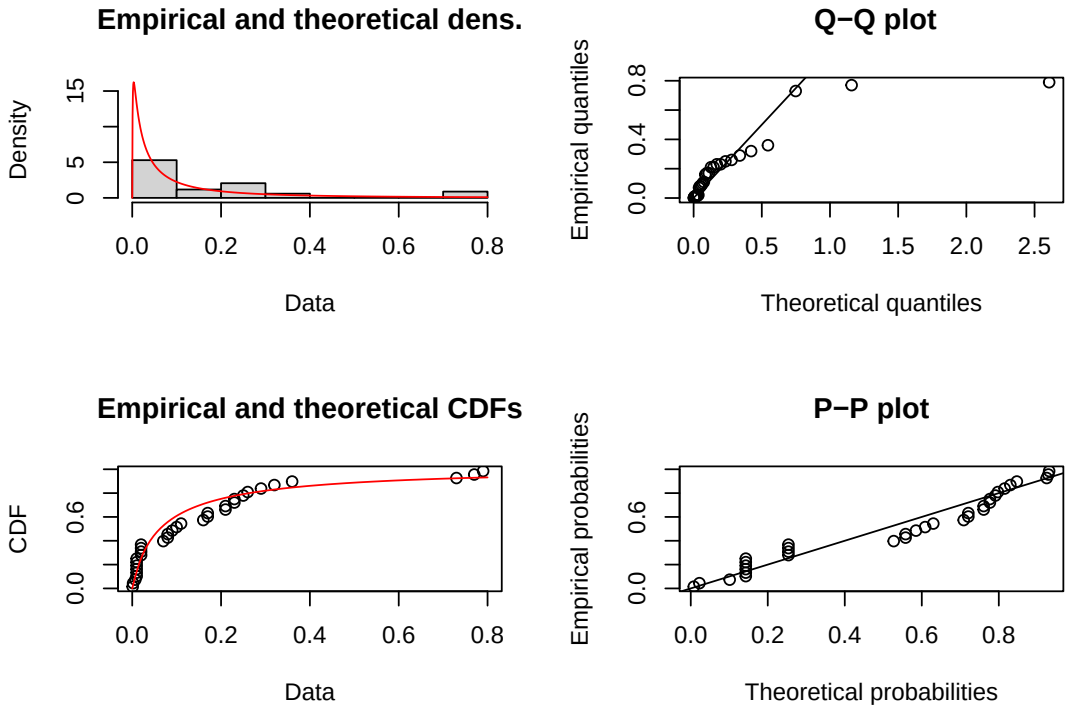

```
## summary statistics
## ------
## min:  0.001   max:  0.79
```

```
## median:  0.095
## mean:  0.1714706
## estimated sd:  0.2148982
## estimated skewness:  1.886472
## estimated kurtosis:  6.23491
```

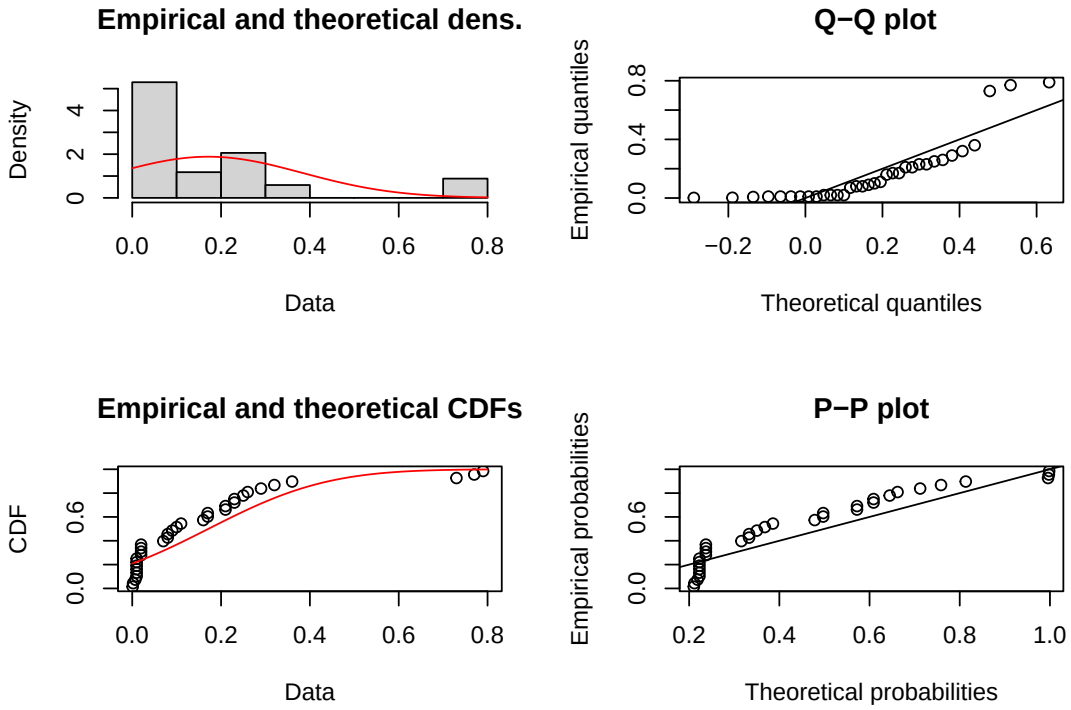
```
## Fitting of the distribution ' beta ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape1 0.5033092  0.1014567
## shape2 2.2686056  0.6220948
## Loglikelihood:  29.19998   AIC:  -54.39995   BIC:  -51.34723
## Correlation matrix:
##           shape1    shape2
## shape1 1.0000000 0.5977919
## shape2 0.5977919 1.0000000
```



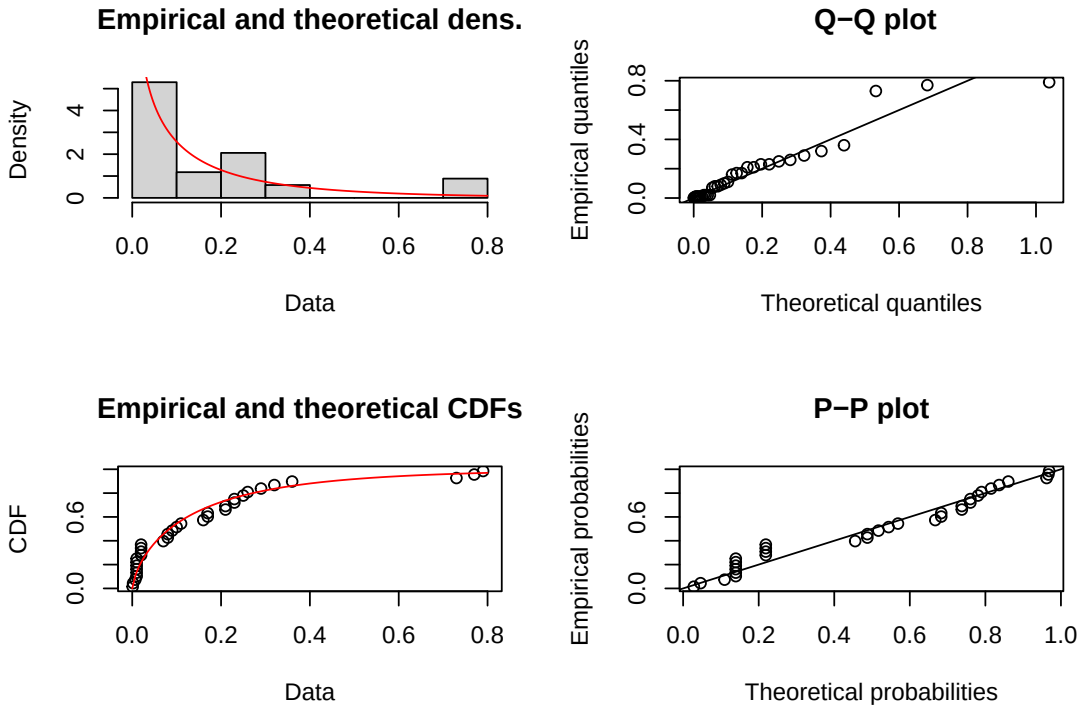

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog -2.775680  0.2940226
## sdlog    1.714432  0.2079051
## Loglikelihood:  27.80042   AIC:  -51.60085   BIC:  -48.54813
## Correlation matrix:
##               meanlog         sdlog
## meanlog  1.000000e+00 -8.143992e-11
## sdlog   -8.143992e-11  1.000000e+00
```




```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 0.1714706 0.03630871
## sd   0.2117143 0.02567156
## Loglikelihood:  4.541683   AIC:  -5.083366   BIC:  -2.030645
## Correlation matrix:
##              mean          sd
## mean 1.00000e+00 8.27872e-13
## sd   8.27872e-13 1.00000e+00
```

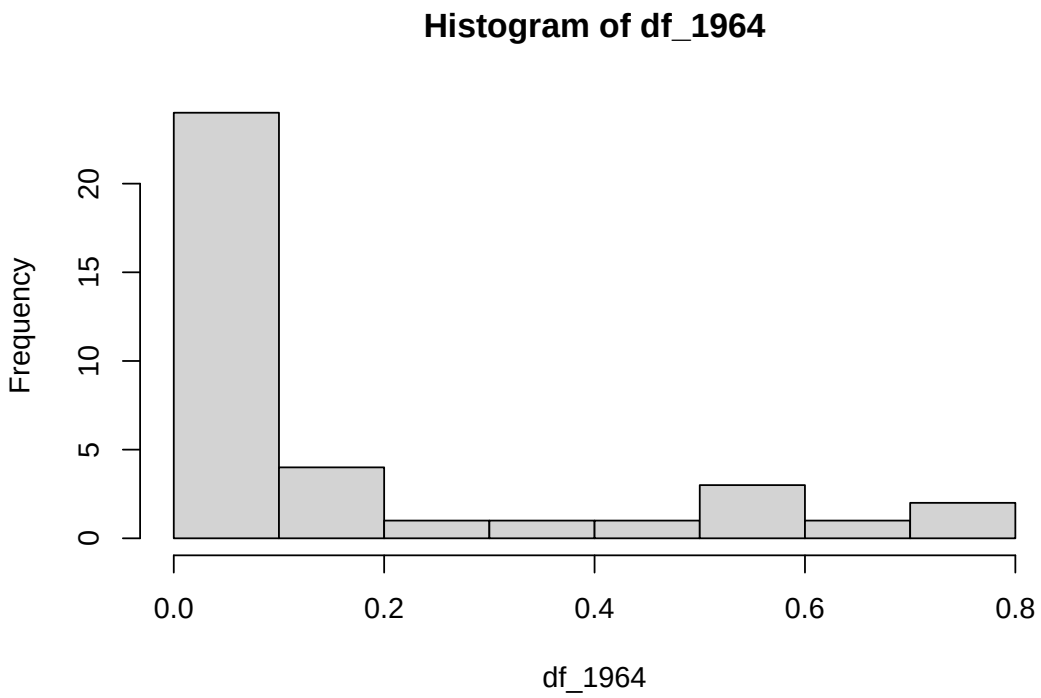


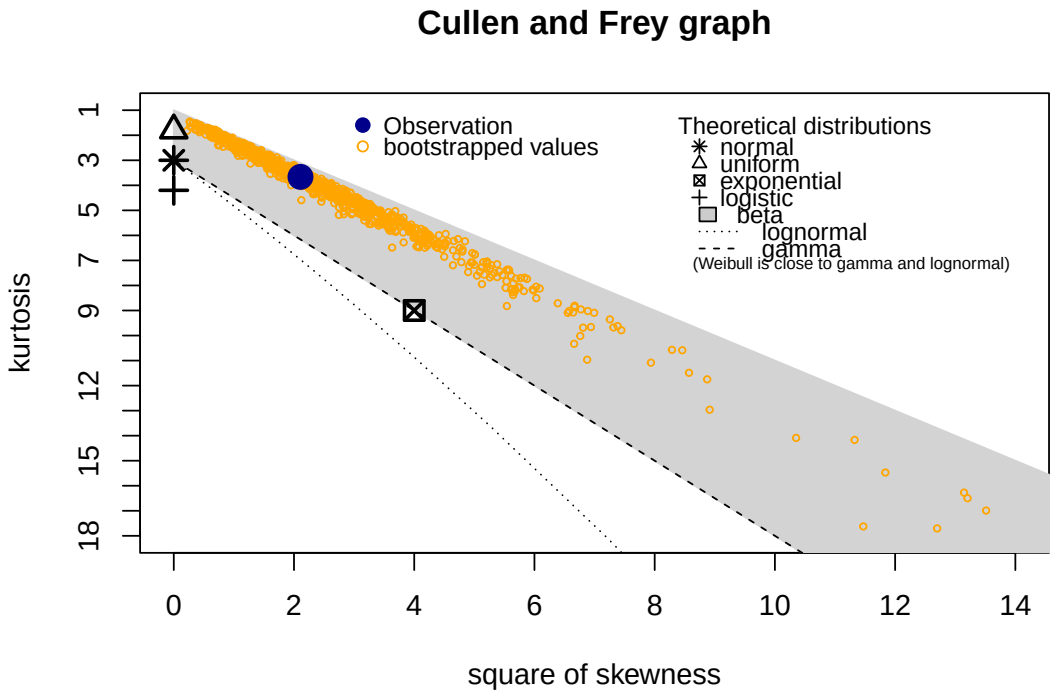

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.7186121 0.09835844
## scale 0.1400436 0.03522616
## Loglikelihood:  29.34843   AIC:  -54.69685   BIC:  -51.64413
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3165628
## scale 0.3165628 1.0000000
```




```
## $distribution
## [1] "Beta"
##
## $sample.size
## [1] 34
##
## $parameters
##    shape1    shape2
## 0.5032853 2.2682753
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mle"
##
## $data.name
## [1] "df_1963"
##
## $bad.obs
## [1] 0
##
## attr(,"class")
## [1] "estimate"
```

The parameters of 1963 rain data are “shape1=0.50” and “shape2=2.27”.

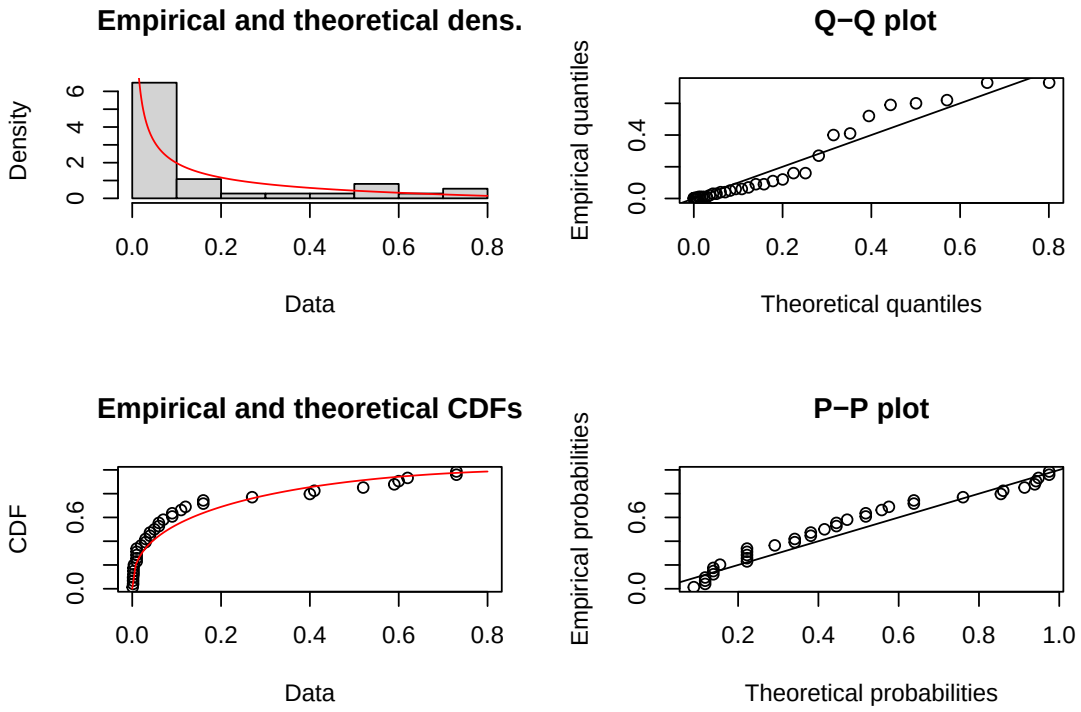




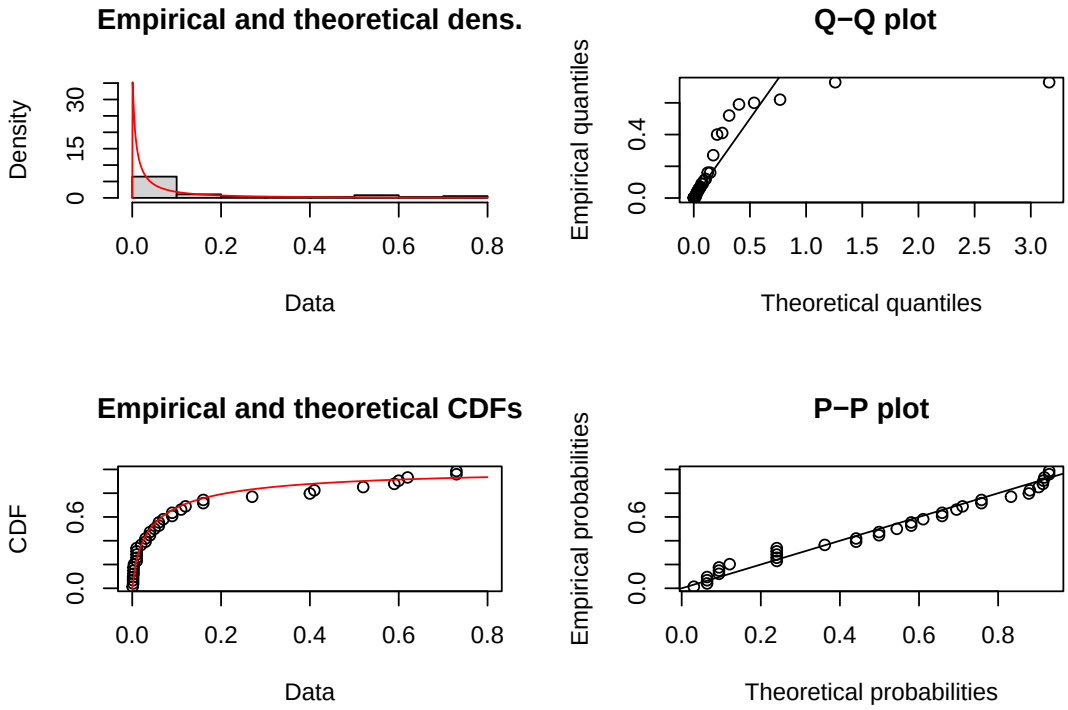

```
## summary statistics
## ------
## min:  0.001   max:  0.73
```

```
## median:  0.05
## mean:  0.1640541
## estimated sd:  0.2326387
## estimated skewness:  1.451831
## estimated kurtosis:  3.66275
```

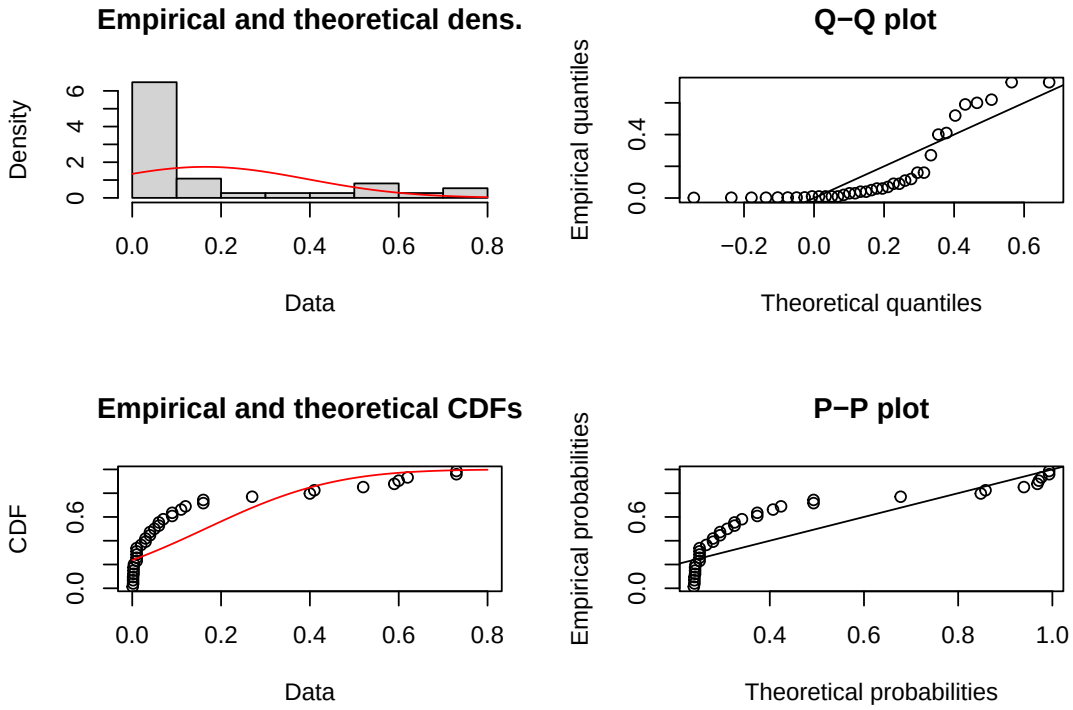
```
## Fitting of the distribution ' beta ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape1 0.3951298 0.07473104
## shape2 1.9347717 0.52783923
## Loglikelihood:  41.1486   AIC:  -78.2972   BIC:  -75.07537
## Correlation matrix:
##           shape1    shape2
## shape1 1.0000000 0.5465862
## shape2 0.5465862 1.0000000
```



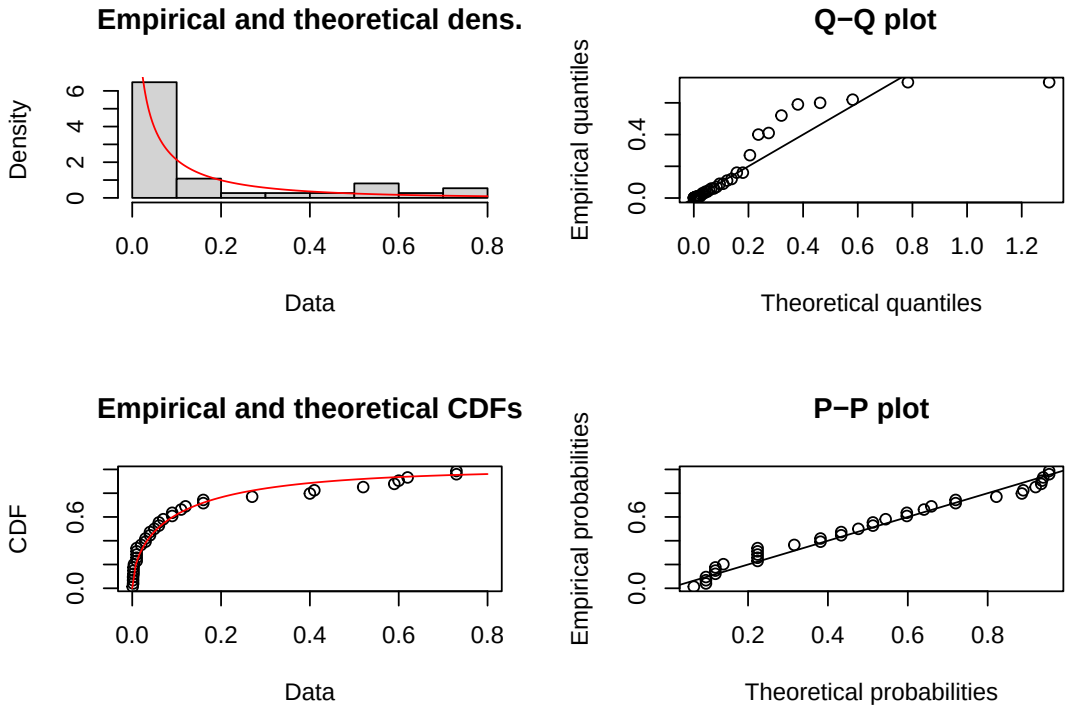

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog -3.213424  0.3245456
## sdlog    1.974134  0.2294881
## Loglikelihood:  41.23116   AIC:  -78.46231   BIC:  -75.24048
## Correlation matrix:
##         meanlog sdlog
## meanlog       1     0
## sdlog         0     1
```




```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 0.1640541 0.03772520
## sd   0.2294734 0.02667346
## Loglikelihood:  1.962093   AIC:  0.0758145   BIC:  3.29765
## Correlation matrix:
##               mean            sd
## mean  1.000000e+00 -3.351525e-13
## sd   -3.351525e-13  1.000000e+00
```




```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.5817657 0.07518619
## scale 0.1057980 0.03160945
## Loglikelihood:  41.03739   AIC:  -78.07478   BIC:  -74.85294
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3250328
## scale 0.3250328 1.0000000
```




```
## $distribution
## [1] "Beta"
##
## $sample.size
## [1] 37
##
## $parameters
##    shape1    shape2
## 0.3951023 1.9345755
##
## $n.param.est
## [1] 2
##
## $method
## [1] "mle"
##
## $data.name
## [1] "df_1964"
##
## $bad.obs
## [1] 0
##
## attr(,"class")
## [1] "estimate"
```

The parameters of 1964 rain data are “shape1=0.40” and “shape2=1.93”. For this analysis, I learned how to identify the distribution of data but I am still trying to learn how to bring all of the years into one model because I found there are different amounts of data sets for each year.So I will try to find a way to analyze it better and then find which year is wet or dry.

In All Likelihood

4.27

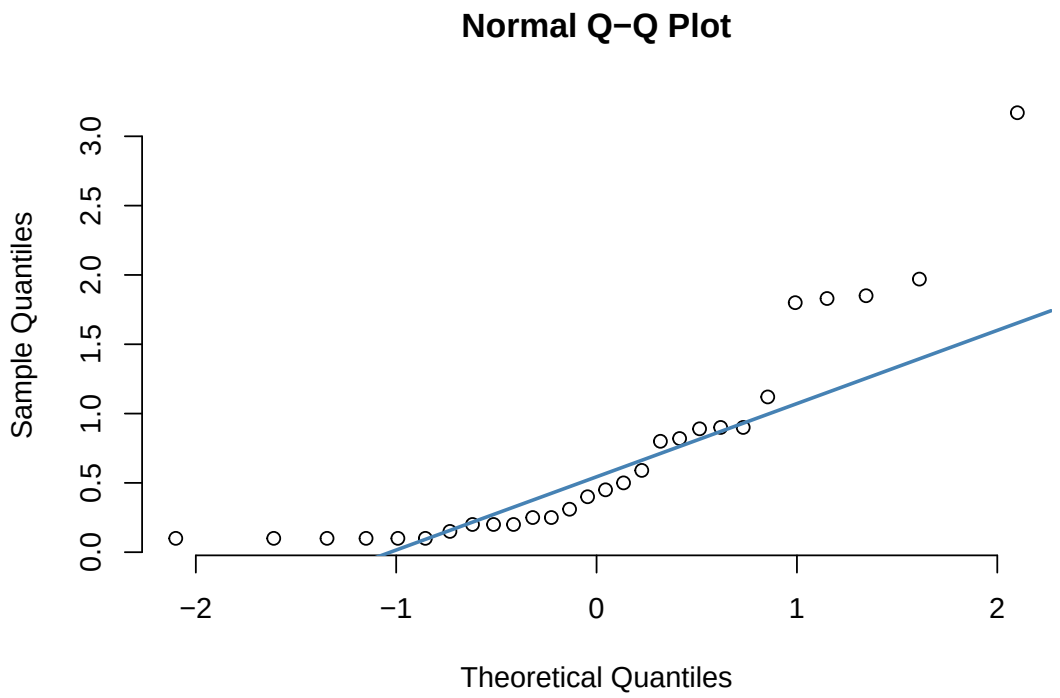
#(a)

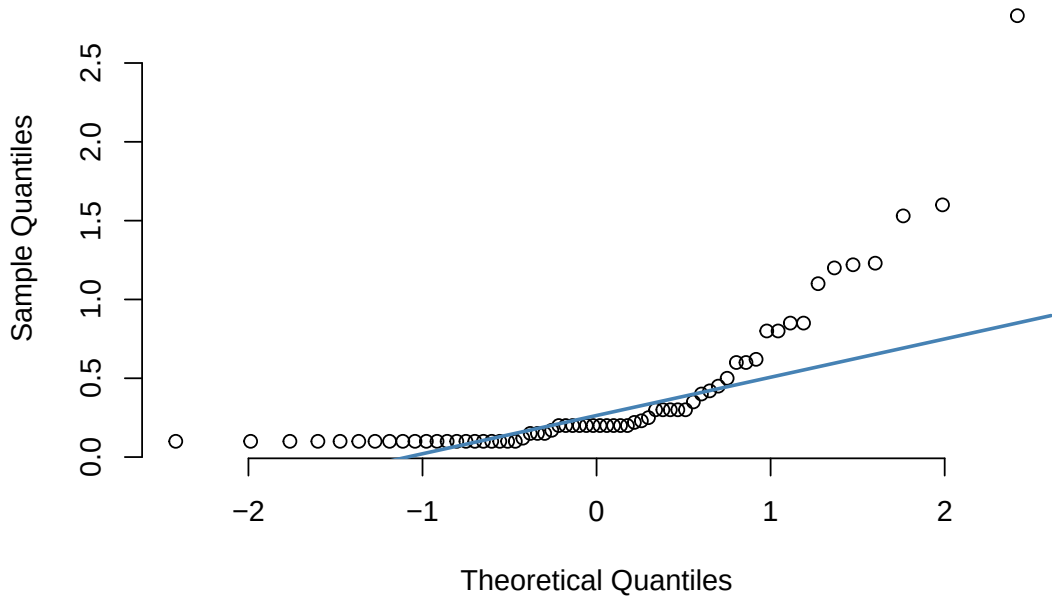
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1875  0.4250  0.7196  0.9000  3.1700
```

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

Here I use summary function to see the descriped statistics of these two data sets. I found for both January and July data,the minimum values are the same – 0.1. The median of January is 0.425 and the median of July is 0.2. The average of January is 0.7196 and that of July is 0.3931. And the maximum values of January and July are 3.17 and 2.8. Generally speaking, there is no much difference between these two data sets.

#(b)





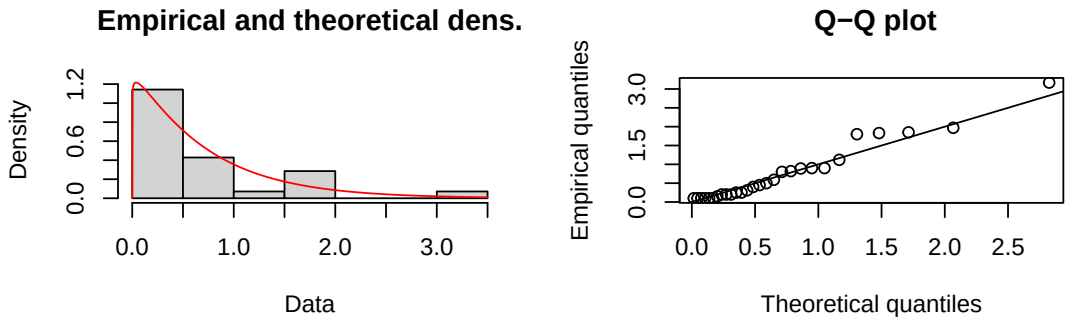
From the QQ-plots, I think these two data sets probably have a gamma distribution.

#(c)

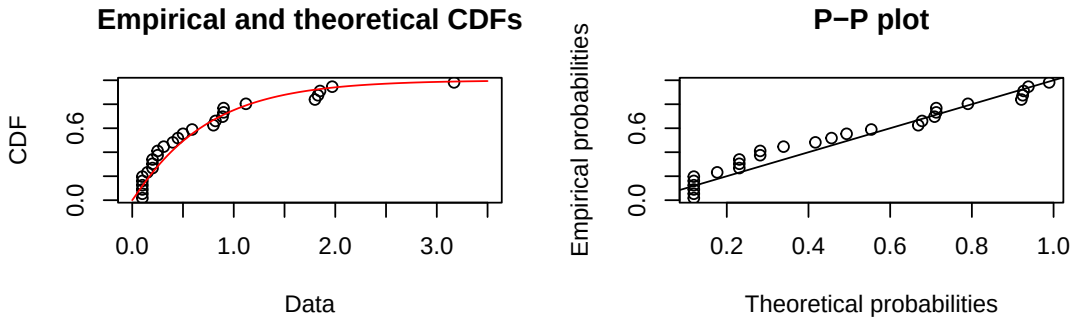
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
## Loglikelihood:  -18.7616   AIC:  41.5232   BIC:  44.18761
## Correlation matrix:
##           shape      rate
## shape 1.0000000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

## 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
```

For the January data, the estimates of parameters are "shape=1.06" and "rate=1.47". The standard errors are 0.25 and 0.44. For the July data, the estimates of parameters are "shape=1.20" and "rate=3.04". The standard error are 0.19 and 0.59.

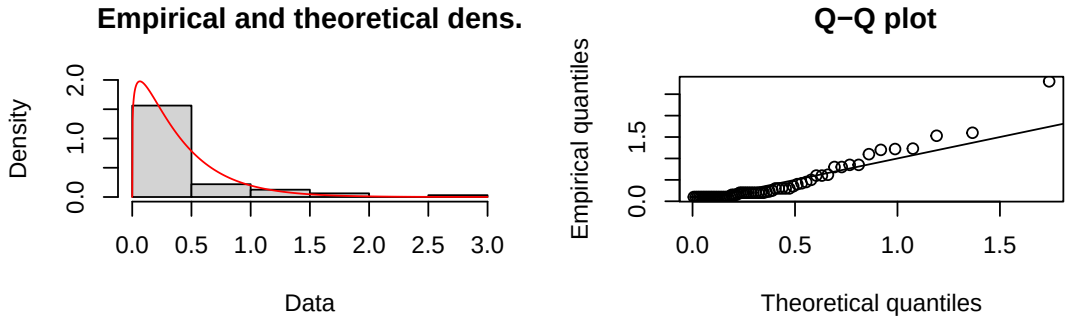
#(d)


Empirical and theoretical dens.
Density
0.0
0.6
1.2
0.0
1.0
2.0
3.0
Data

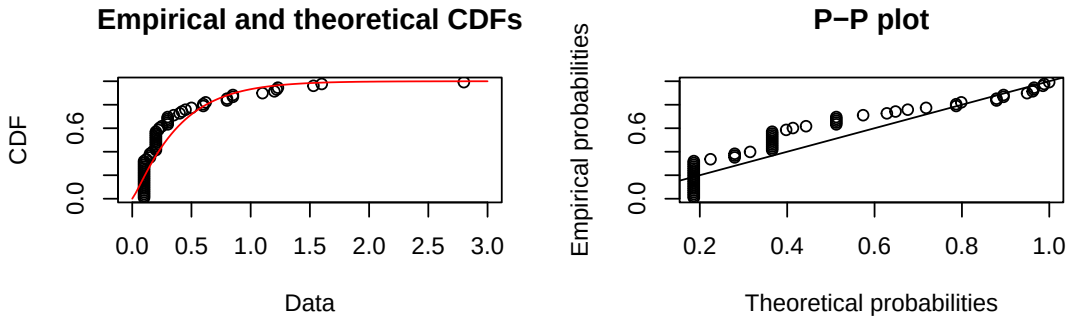
Q−Q plot
Empirical quantiles
0.0
1.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
Theoretical quantiles


Empirical and theoretical CDFs
CDF
0.0
0.6
0.0
1.0
2.0
3.0
Data

P−P plot
Empirical probabilities
0.0
0.6
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities


Empirical and theoretical dens.
Density
0.0
1.0
2.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Data

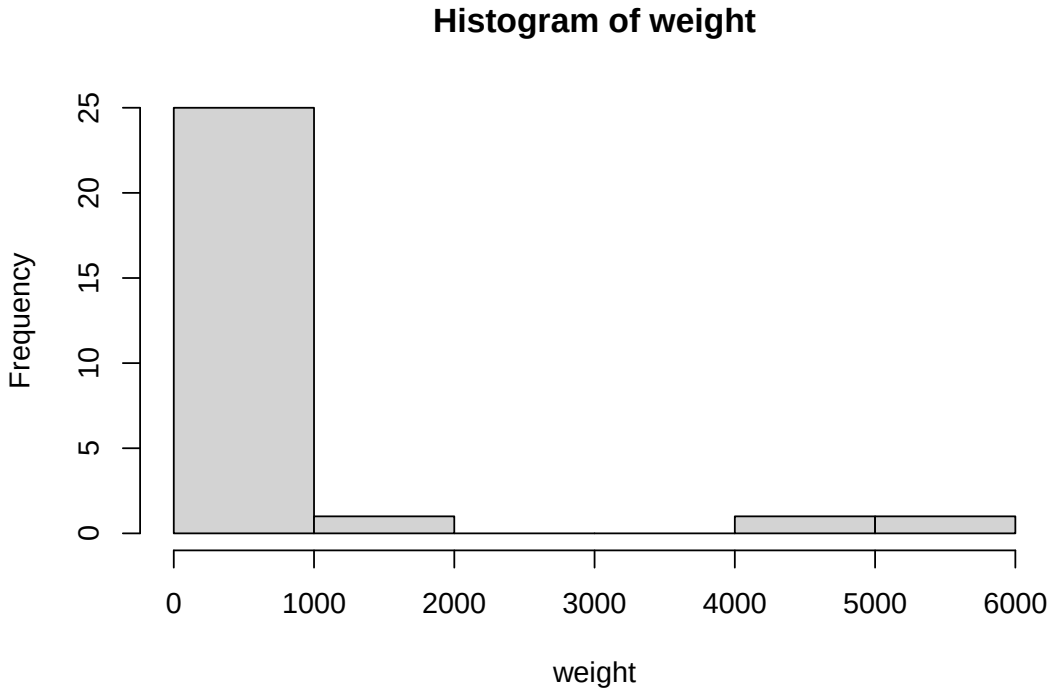
Q−Q plot
Empirical quantiles
0.0
1.5
0.0
0.5
1.0
1.5
Theoretical quantiles


Empirical and theoretical CDFs
CDF
0.0
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Data

P−P plot
Empirical probabilities
0.0
0.6
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities

4.39

When using R,we can make use of the boxcox function from the MASS package to estimate the transformation parameter by maximum likelihood estimation.This function will also give us the 95% confidence interval of the parameter.

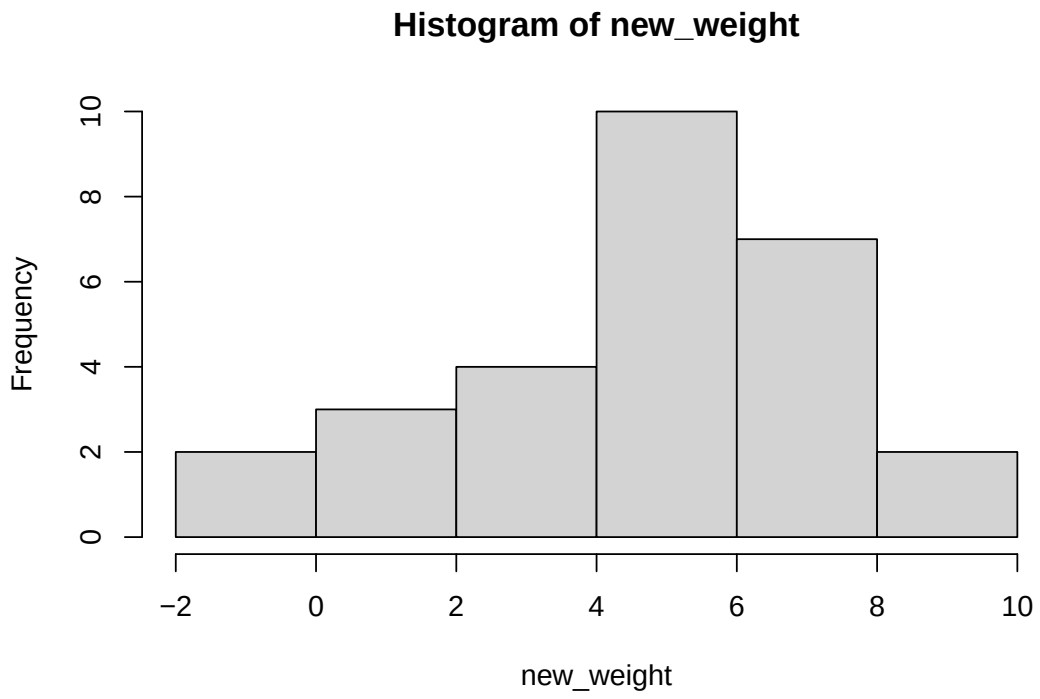


In order to calculate the optimal lamdba I have to compute a linear model with the lm function and pass it to the boxcox function as follows.

```
## $lambda
## [1] -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0
##
## $objective
## [1] 0.4720085 0.5201090 0.6102865 0.7790260 0.9801890 0.8488209 0.6592582
## [8] 0.5776345 0.5443564
##
## $objective.name
## [1] "PPCC"
##
## $optimize
## [1] FALSE
##
## $optimize.bounds
## lower upper
##    NA    NA
##
## $eps
## [1] 2.220446e-16
```

```
##
## $lm.obj
##
## Call:
## lm(formula = weight ~ 1, y = TRUE, qr = TRUE)
##
## Coefficients:
## (Intercept)
##       574.5
##
##
## $sample.size
## [1] 28
##
## $data.name
## [1] "lm(weight ~ 1)"
##
## attr(,"class")
## [1] "boxcoxLm"
```

Note that the center objective is 0.98 which represents the estimated parameter lambda and the others the 95% confidence interval of the estimation.As the previous shows that the 0 is inside the confidence interval of the optimal lambda and as the estimation of the parameter is really close to 0 in this question, the best option is to apply the logarithmic transformation of the data.



Now the data looks more like following a normal distribution, but I also use a statistical test to check it, as the Shapiro-Wilk test.

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

As the p-value is greater than the usual levels of significance (1%,5% and 10%) we have no evidence to reject the null hypothesis of normality.