MA 677 Final

Jack Carbaugh

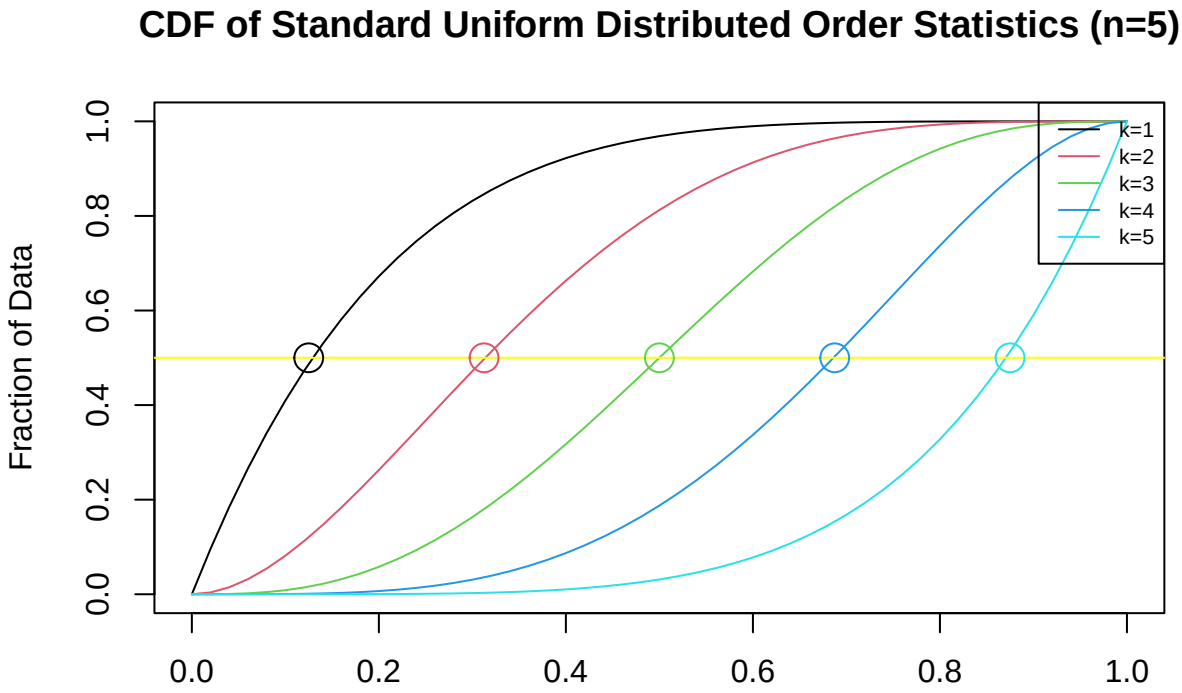
5/7/2022

Planning

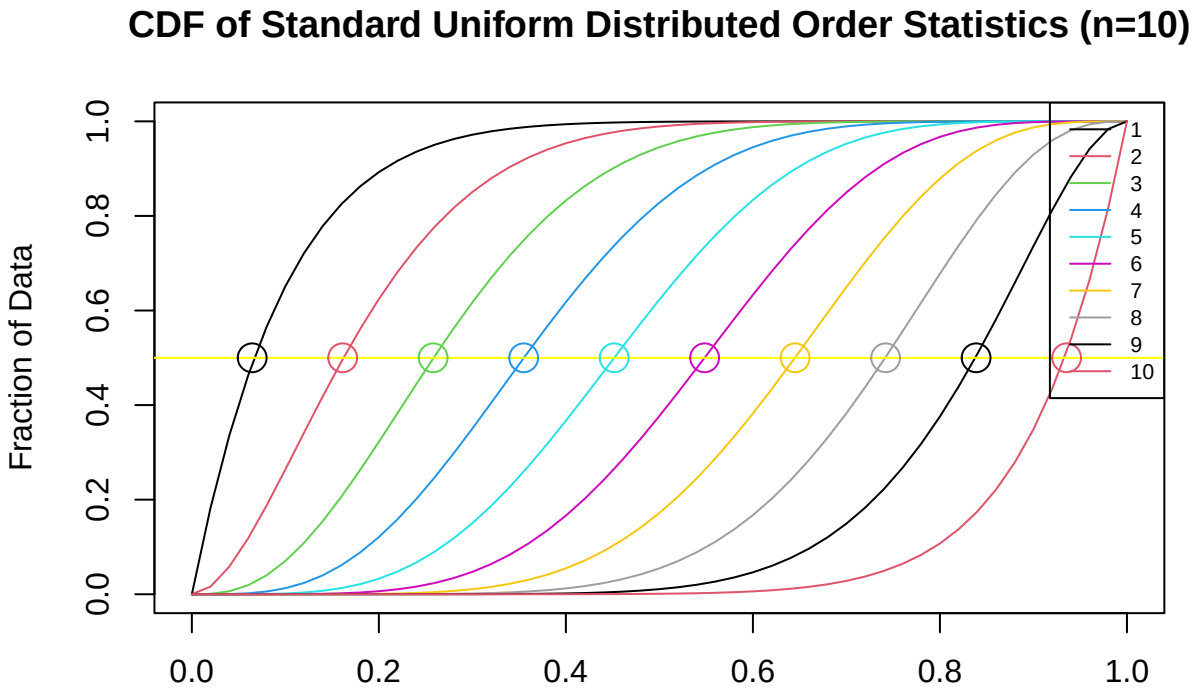
The planning for this project began with reading the relevant sections of Pawitan's In All Likelihood, in order to better understand the example problems. Once completed, the assigned exercises would be attempted. I would mainly follow Pawitan's work, and observe similar examples if I got stuck. It was also helpful to search for R packages to directly calculate some of the tasks. After the exercises, I planned to use the previous examples, mainly 4.27, to begin the Illinois rain analysis. The rain analysis required learning about some distributions, and other common facts about hydrology. Finally, the results would be discussed to think about generalizations and other applications.

Pawitan Exercise 4.25

From exercise 2.4, it was shown that the kth order statistic from a U(0,1) distribution has a distribution of Beta(k, n-k+1).

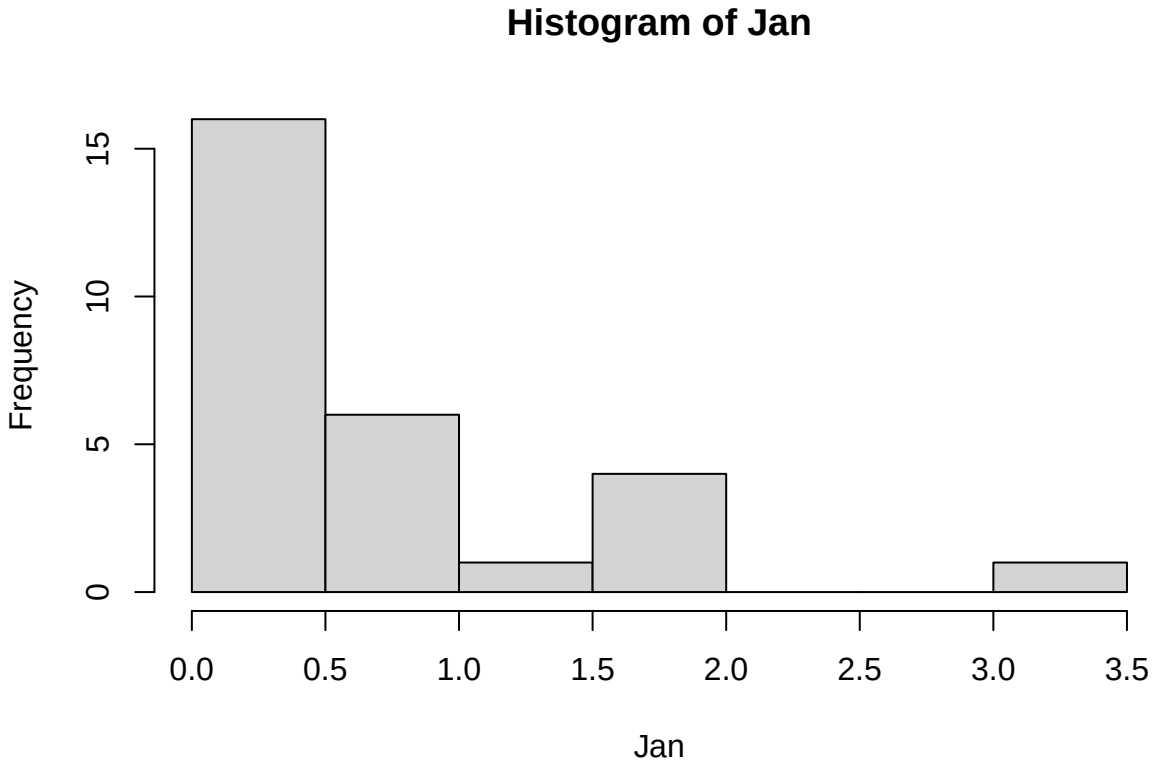


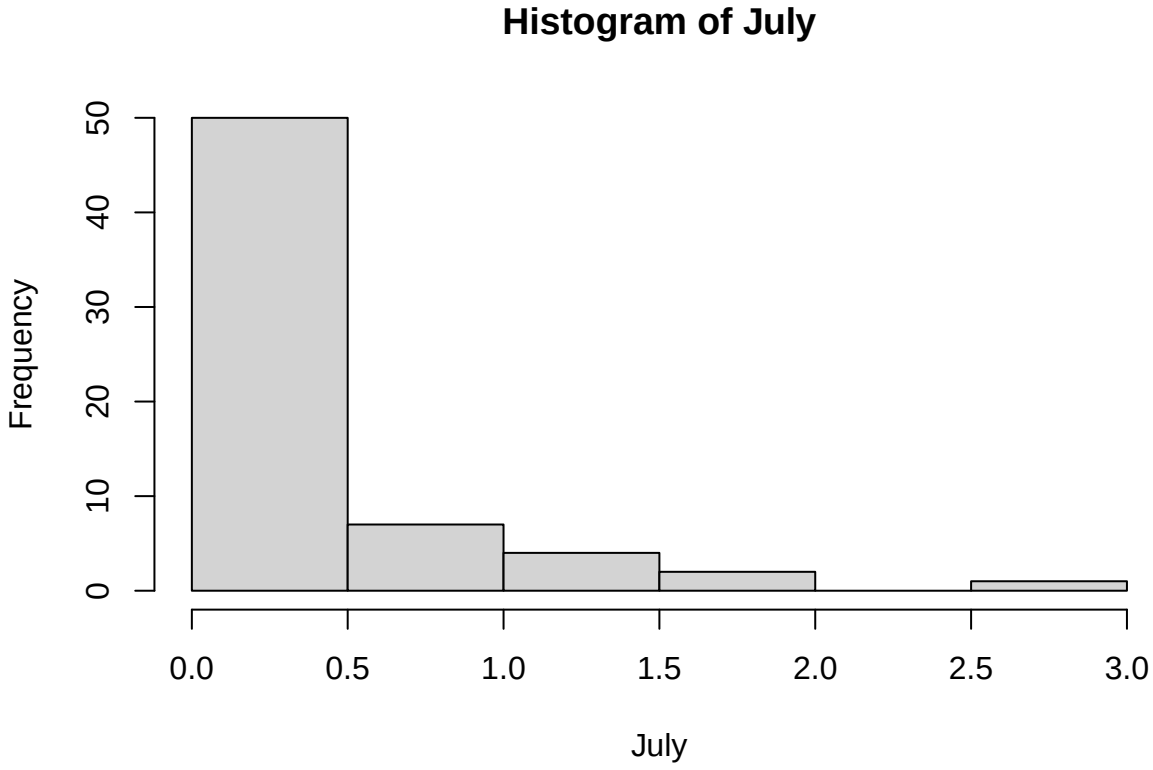
The above plot shows the CDFs of the 5 order statistics, based on their appropriate beta distributions. The yellow line is at 0.5, and thus shows where the median of each order statistic would be. The colored dots are calculated based on the approximation $median(U_{(i)}) \approx \frac{i-1/3}{n+1/3}$. As can be seen, the approximated medians match up almost exactly with their actual values, with the more extreme values (1 and 5) having slightly more error.



The above plot shows the exact same concept as before, now with n=10 for the order statistics. Once again, we see that the approximated medians match the actual medians very well.

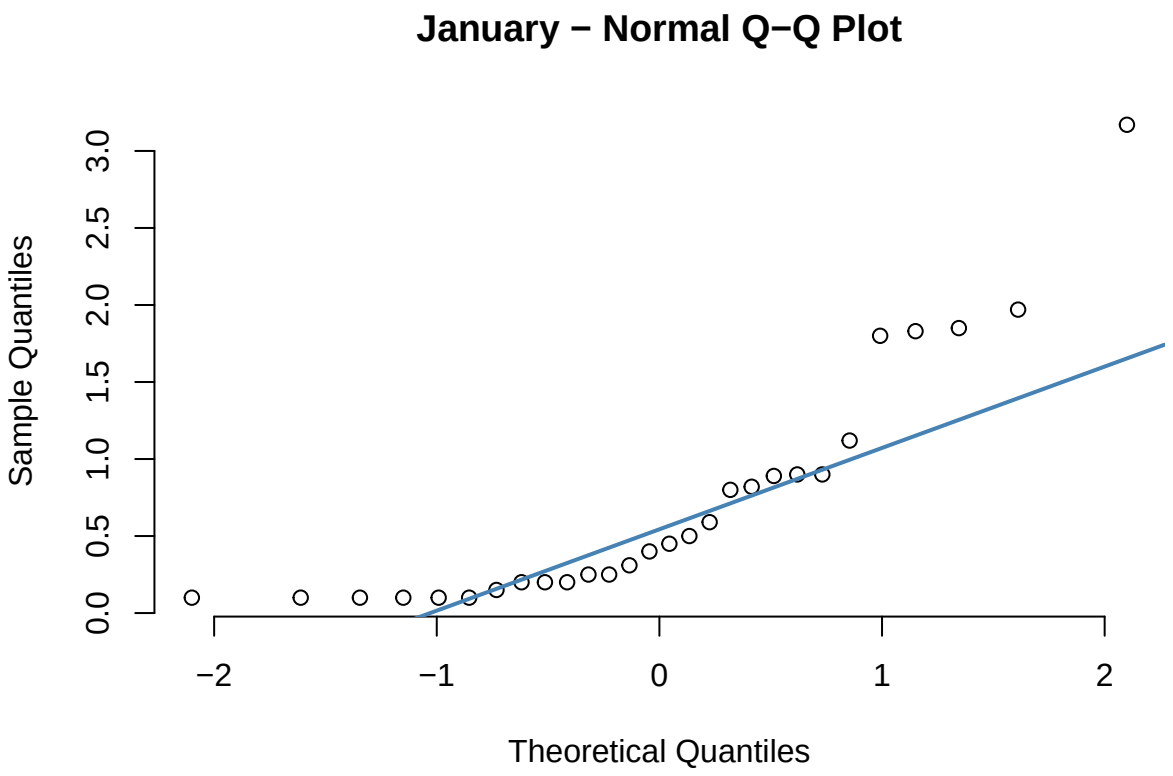
Pawitan Exercise 4.27

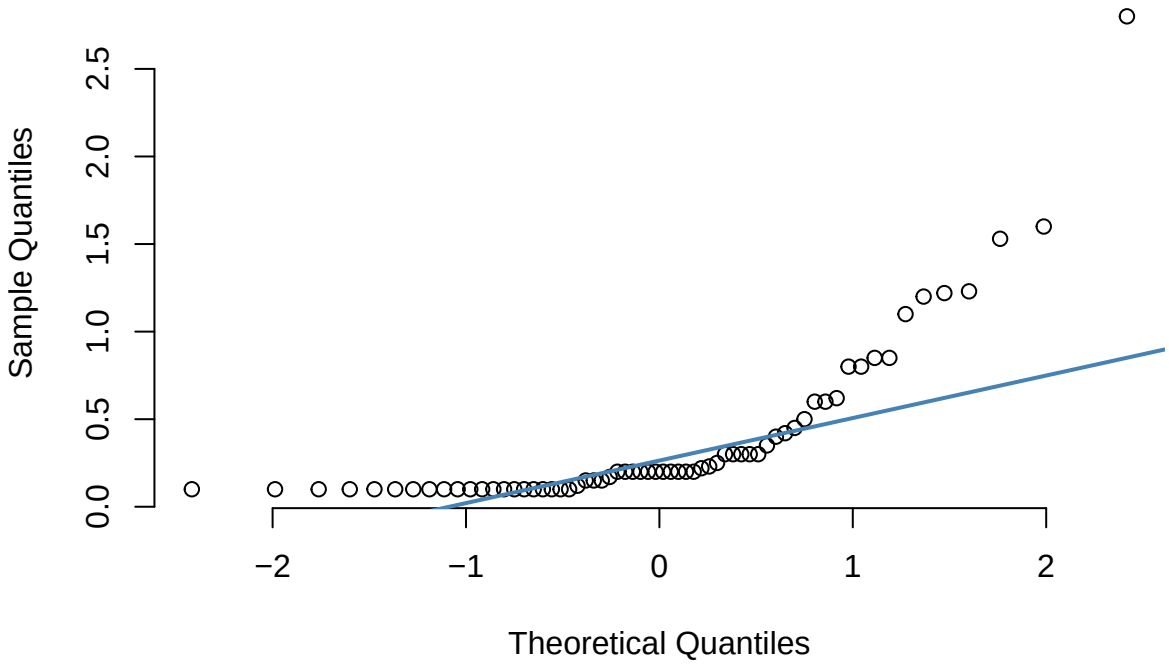




```
## Mean January Rainfall: 0.72

## Standard deviation of January Rainfall: 0.77

## Skewness of January Rainfall: 1.47

## Mean July Rainfall: 0.39

## Standard deviation of July Rainfall, 0.48

## Skewness of July Rainfall: 2.64
```

Looking at the summary statistics, we see January had more rainfall per storm on average than July. The rainfall of January's storms was more variable than July's, while the rainfall of July's storms was more right skewed than January's.

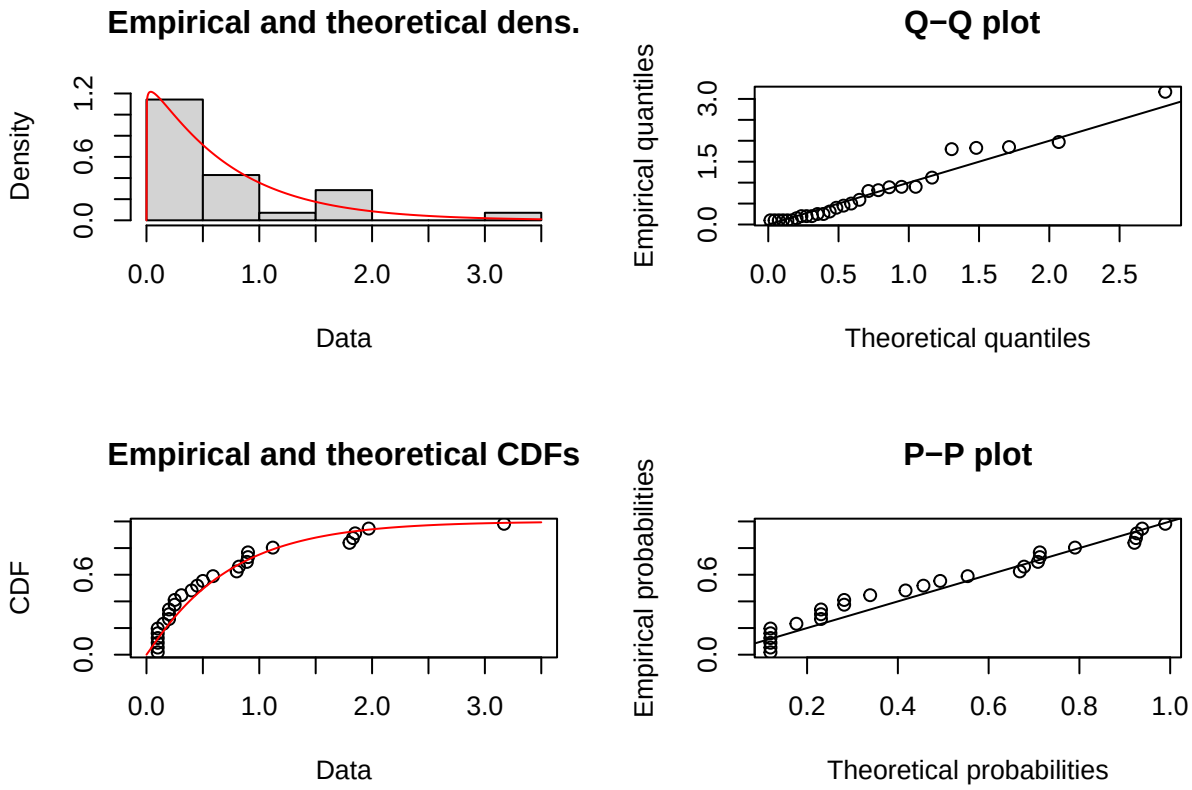




Both QQ Plots show a lot of extreme values on the tails. This amount of skewness is characteristic of a right-skewed exponential distribution. Thus, an exponential distribution would be a reasonable for this data.

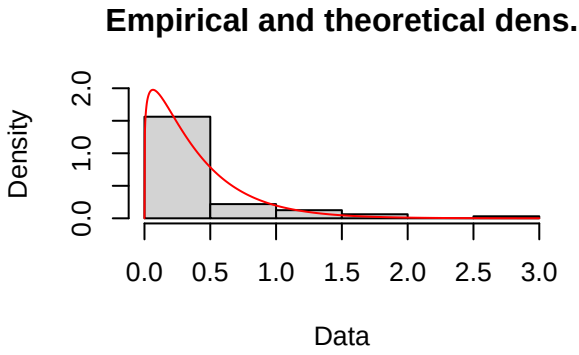
Fitting a gamma distribution to the January data:

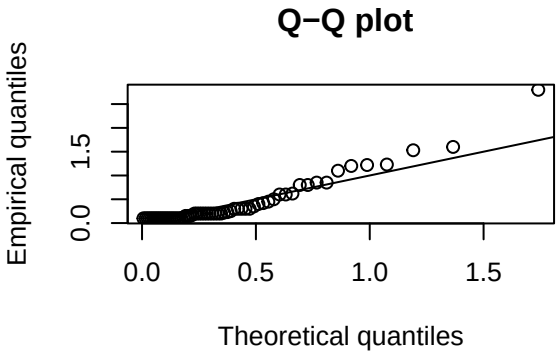
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
## Loglikelihood:  -18.7616   AIC:  41.5232   BIC:  44.18761
## Correlation matrix:
##           shape      rate
## shape 1.0000000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

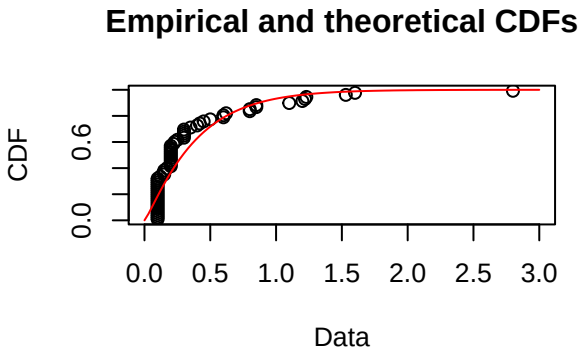


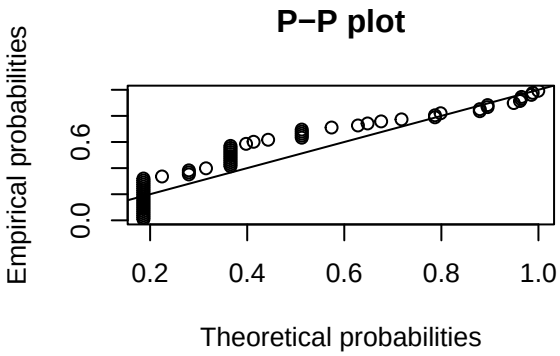
Fitting a gamma distribution to the July data:

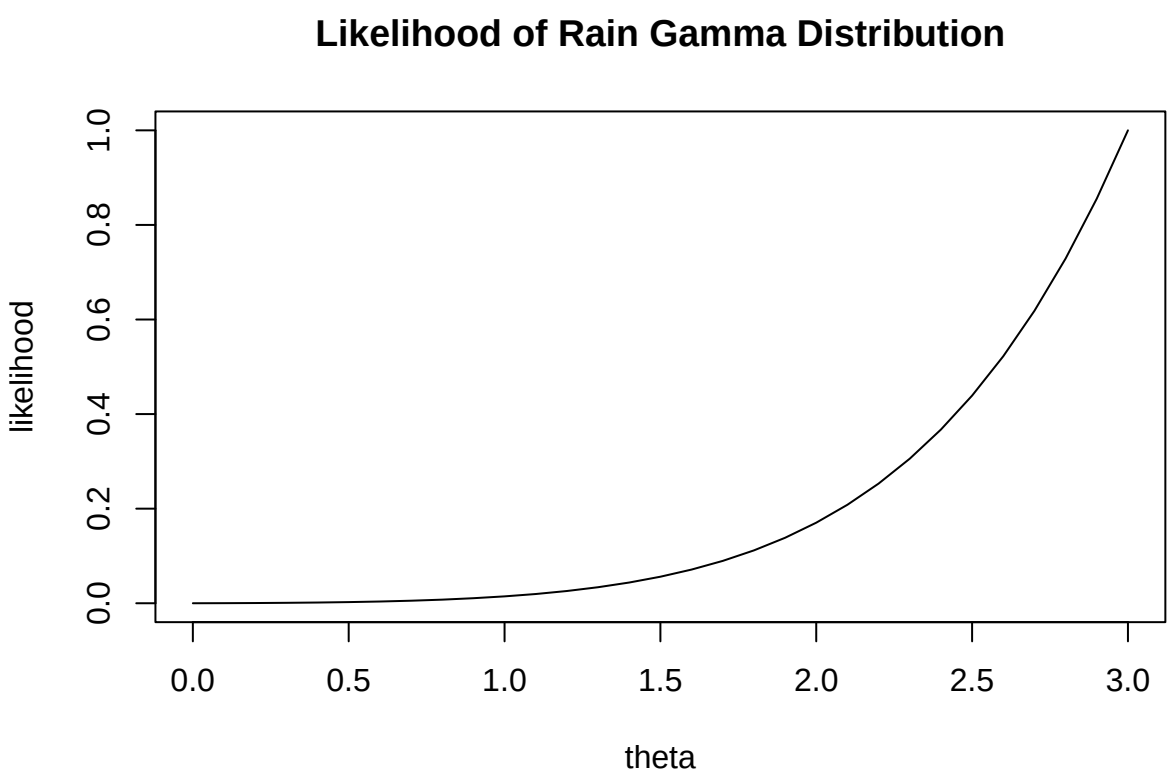
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
## Loglikelihood:  -3.634886   AIC:  11.26977   BIC:  15.58754
## Correlation matrix:
##           shape      rate
## shape 1.0000000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


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

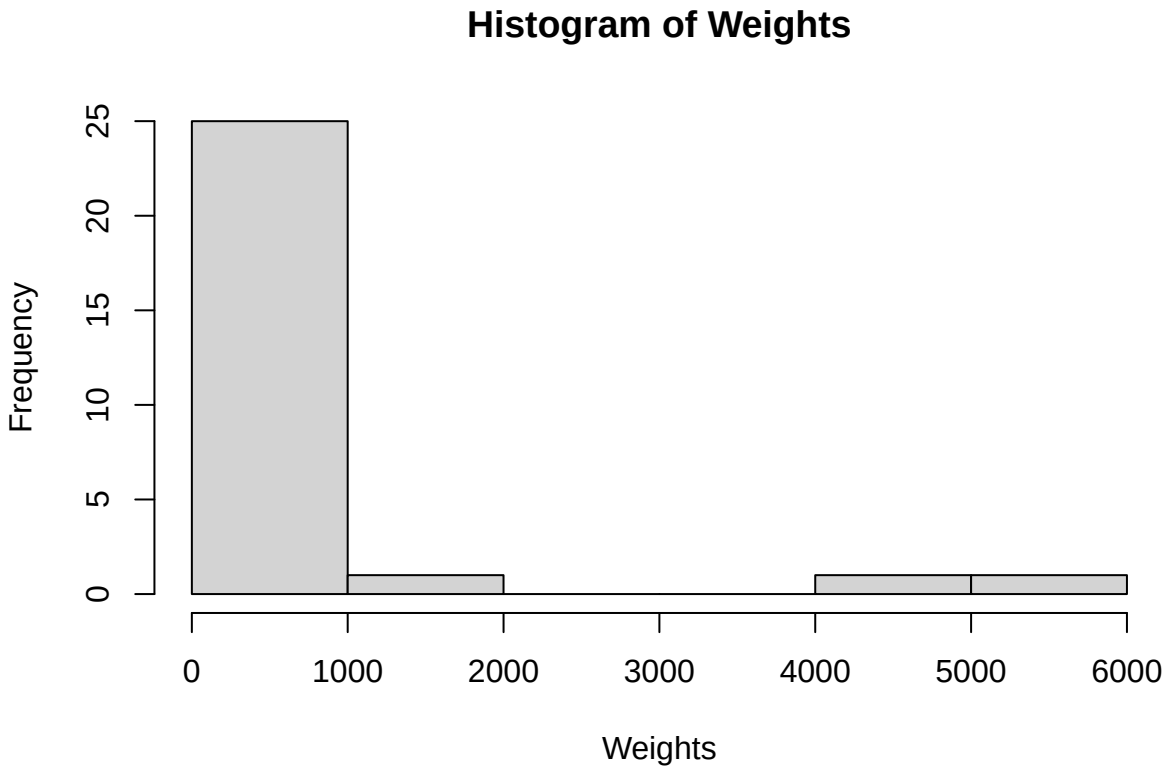

Empirical and theoretical CDFs
CDF
0.0
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Data

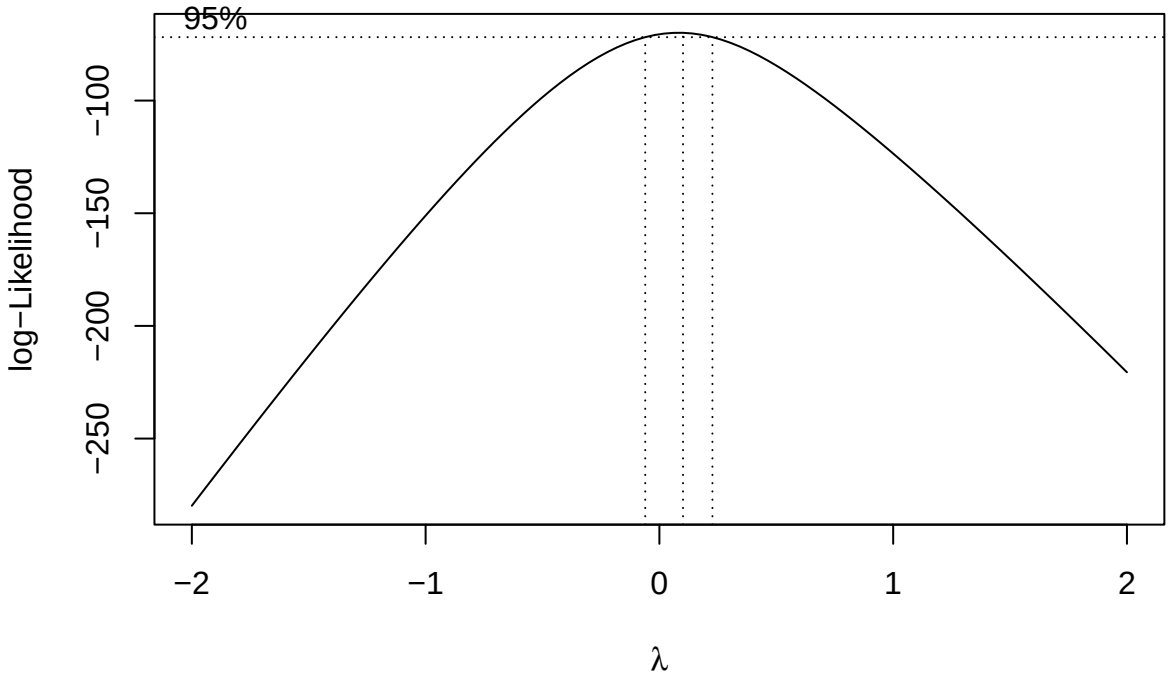

P−P plot
Empirical probabilities
0.0
0.6
0.2
0.4
0.6
0.8
1.0
Theoretical probabilities

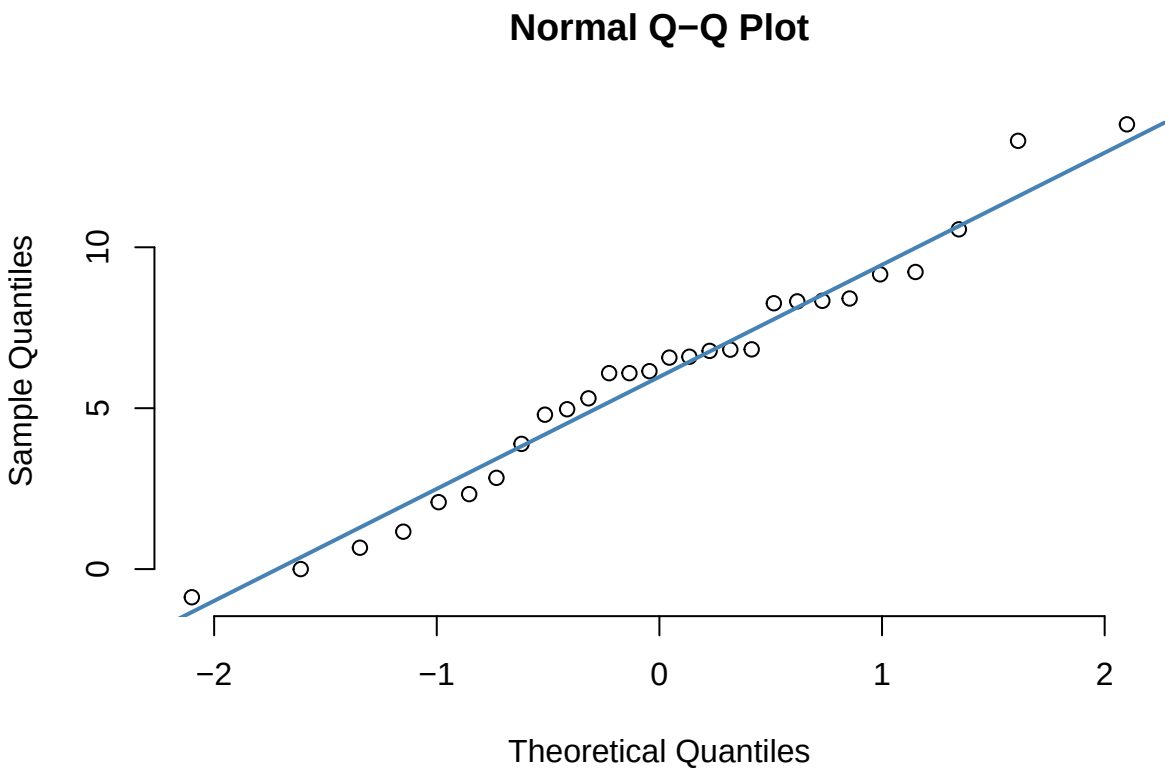


Overall, according to the QQ plot, the gamma models seem to work well for both months of the rainstorm data. We see that the two months have similar estimated shape parameters, at 1.05 for January and 1.19 for July. July's estimated rate parameter of 3.04 is higher than January's estimated rate of 1.46, explaining how July's data was less spread.

Pawitan Exercise 4.39

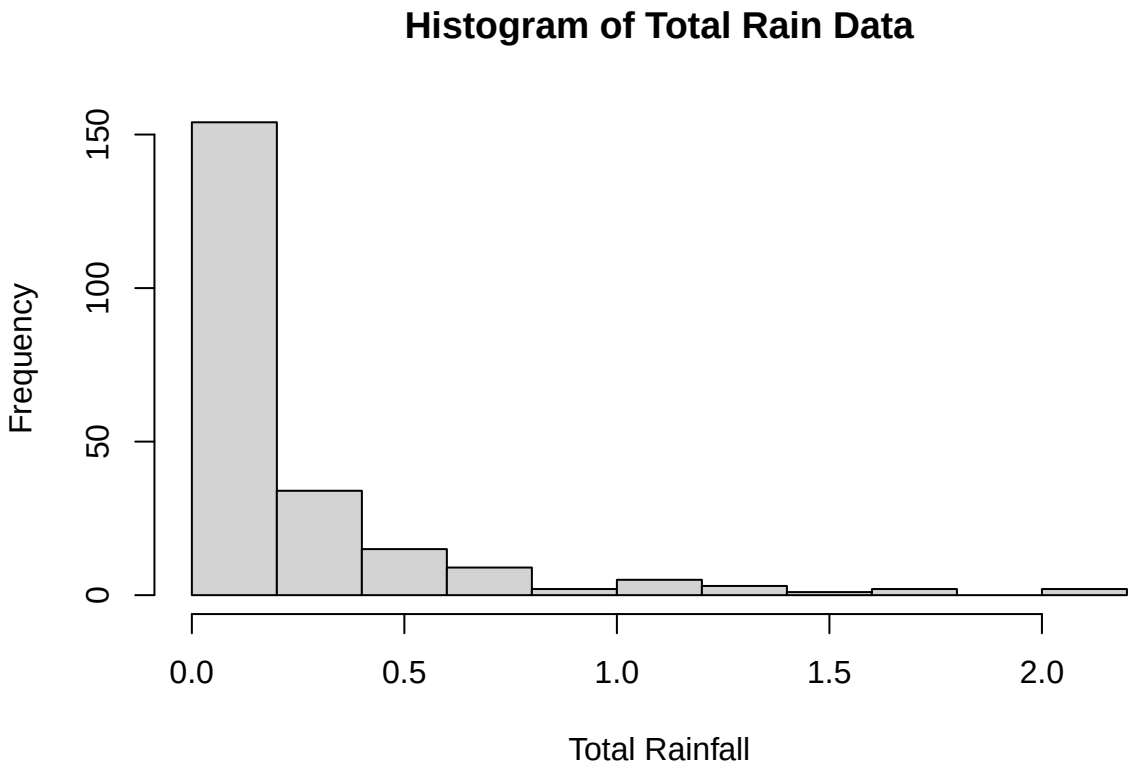






The box-cox transformation showed that a lambda value of about $\frac{10}{99}$ would normalize the dataset. After the transformation, the QQ plot of the data shows the points hugging the normal line, meaning the transformation appears to be successful.

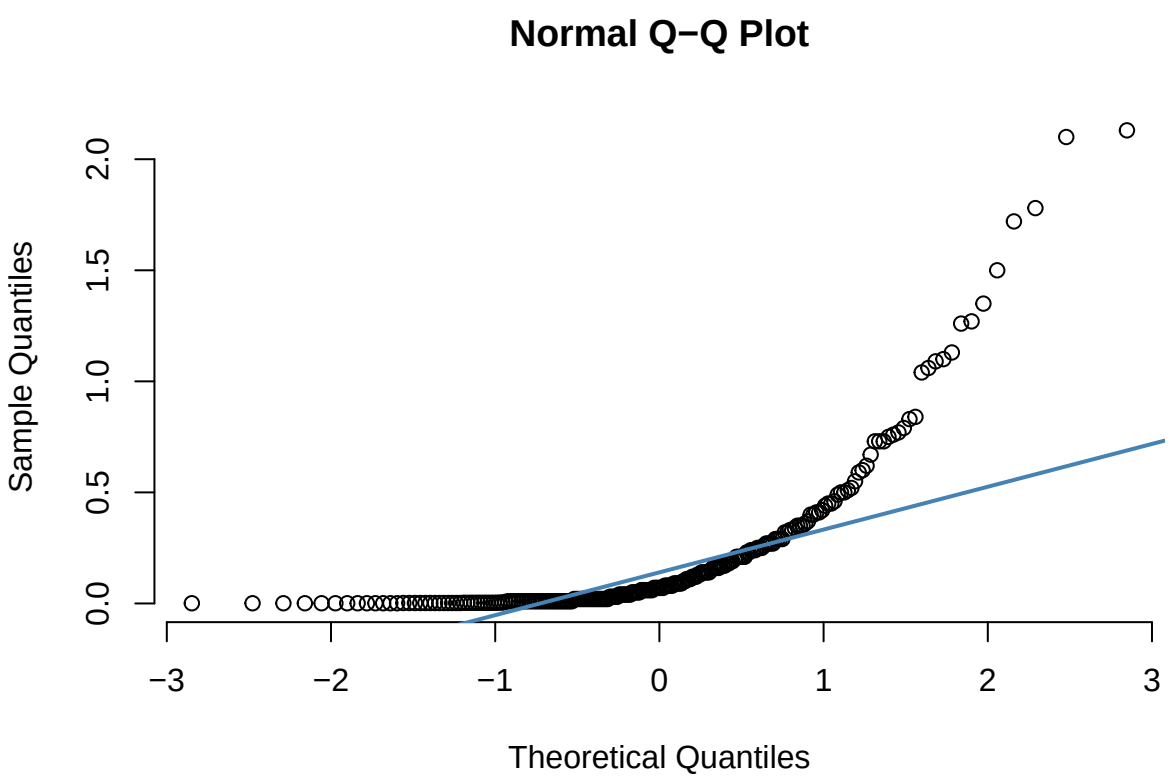
Illinois Rainfall Analysis




```
## Mean Rainfall: 0.22
```

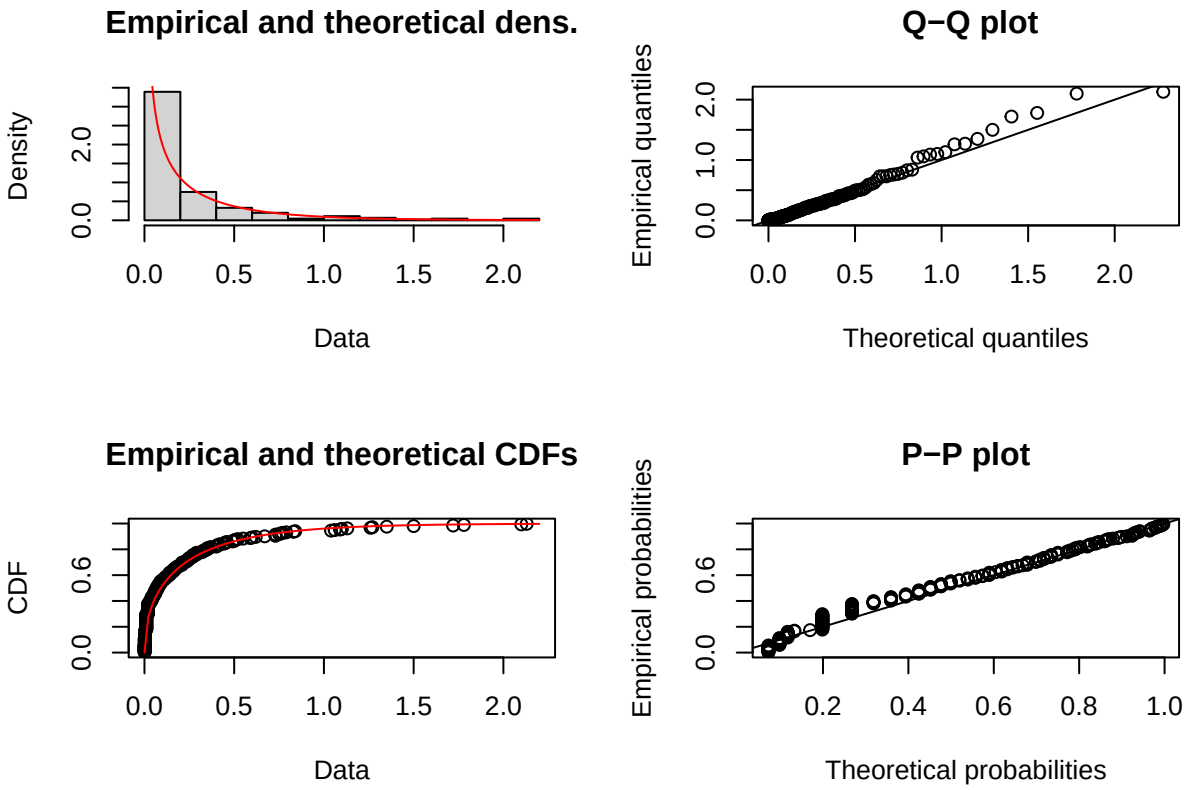
```
## Standard deviation of Rainfall: 0.37
```

```
## Skewness of Rainfall: 2.74
```

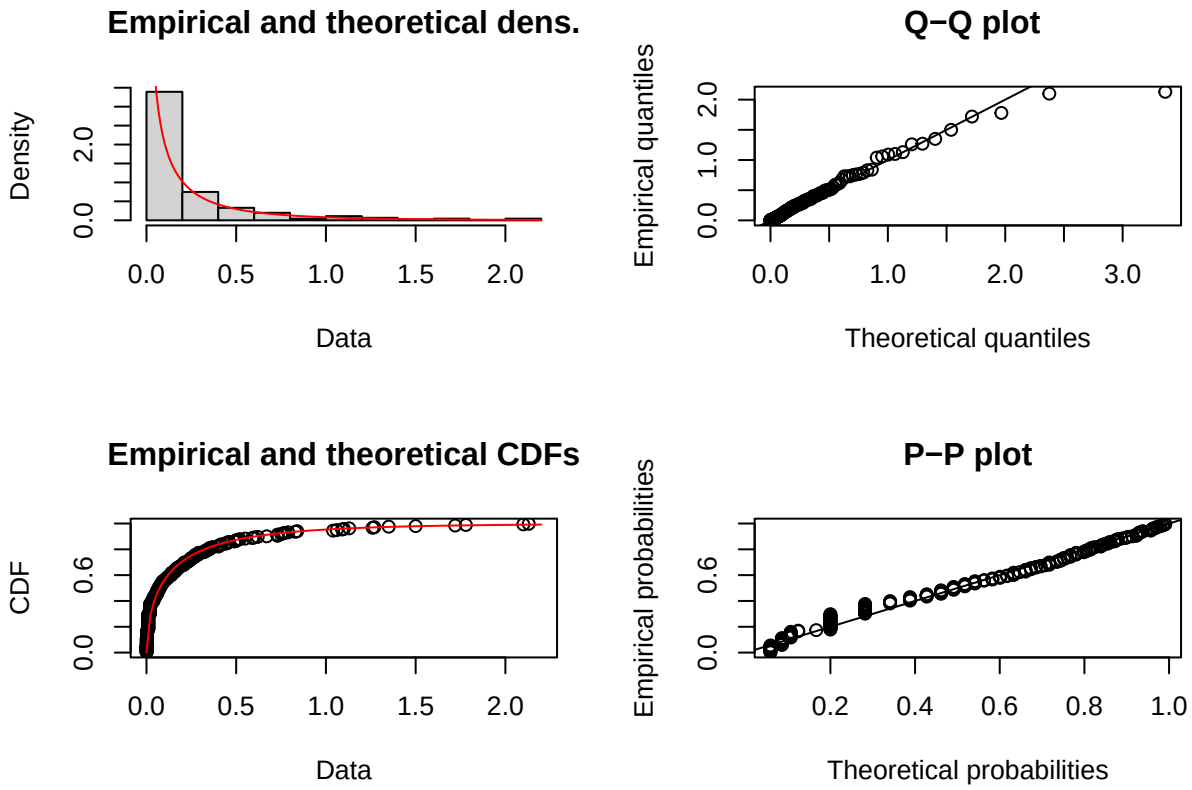


Looking at the histograms of the entire dataset and the individual years, some form of exponential distribution would appear to fit the data well. The summary statistics show a mean near 0, and a high right skewness. This right skewness is confirmed by the QQ plot, and also signals that an exponential distribution may fit this data. Some exponential distributions we can check are the gamma, weibull, and lnorm.

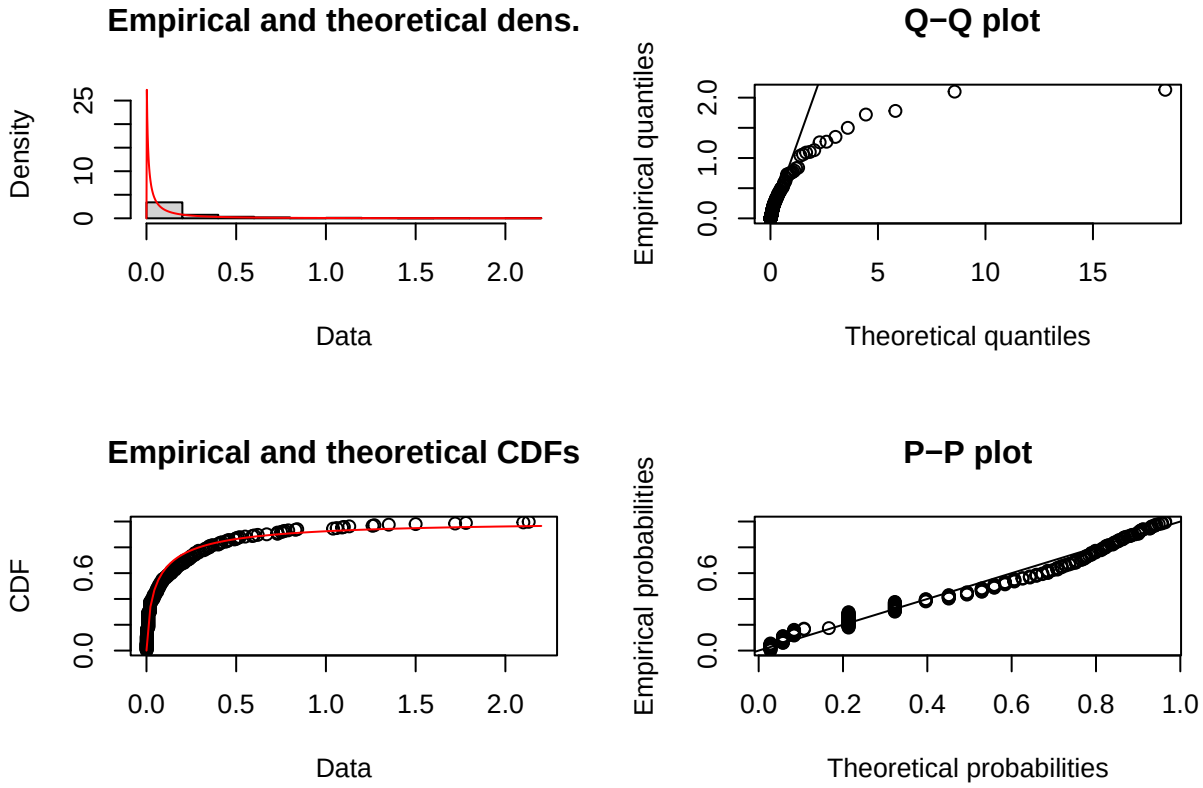
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.4408386  0.0337663
## rate  1.9648409  0.2474440
## Loglikelihood:  185.3477   AIC:  -366.6954   BIC:  -359.8455
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.6082109
## rate  0.6082109 1.0000000
```




```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.5690986 0.02959323
## scale 0.1394868 0.01718016
## Loglikelihood:  189.4436   AIC:  -374.8873   BIC:  -368.0374
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3219421
## scale 0.3219421 1.0000000
```




```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## meanlog -2.964185 0.13692665
## sdlog    2.063008 0.09682166
## Loglikelihood:  186.3855   AIC:  -368.7709   BIC:  -361.921
## Correlation matrix:
##         meanlog sdlog
## meanlog       1     0
## sdlog         0     1
```



For right skewed, exponential data such as this, some potential distributions for the data could be the gamma, weibull, and lnorm. Above, we attempt to fit these 3 using the fitdistrplus package and maximum likelihood estimation for the parameter estimates. From the density and QQ plots, it appears as though the gamma and weibull distributions are potential fits, while the lnorm is not.

Deciding between the gamma and weibull distributions is a bit tricky. This is especially true since the weibull distribution is already related, being a special case of a generalized gamma distribution, and has similar parameters (shape and scale/rate). The gamma distribution is more well known the weibull distribution, and both densitys appear to be very similar. However, the QQ plot for the weibull is noticeably more linear for the upper values than the gamma, and as well the theoretical cdf matches more closely to the empirical for the lower values. A potential problem for the weibull distribution is the highest value (2.13 inches) is a noticeable outlier, while is not so much for the gamma.

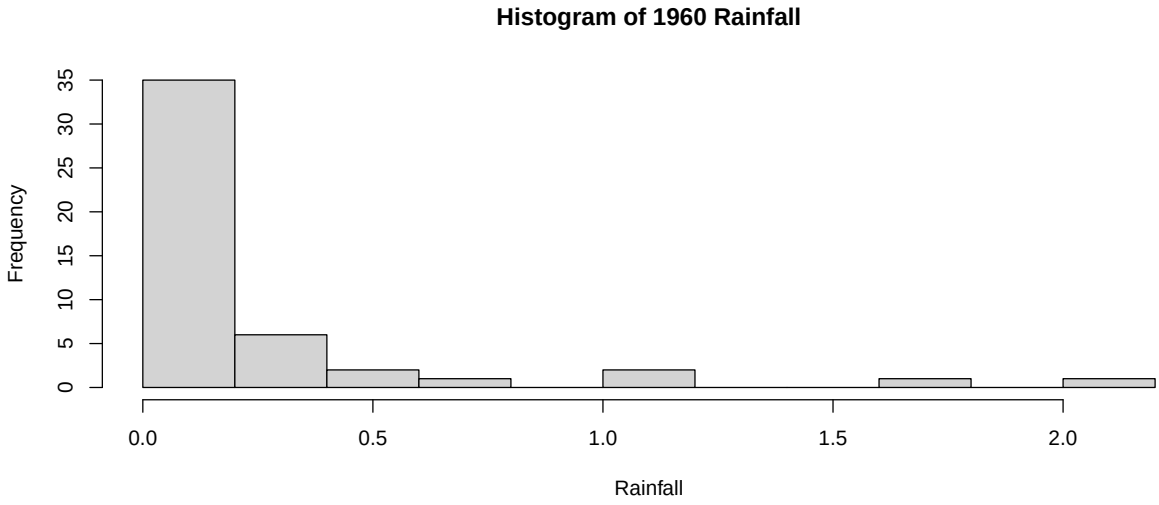
Both distributions appear to fit the Illinois rain data very well, and both are known to model daily rain data well. The gamma distribution was shown to work well for similar data in Pawitan exercise 4.27, and the weibull distribution's usefulness in hydrology has been discussed by Singh (https://link.springer.com/chapter/10.1007/978-94-017-1431-0_12). As both distributions are well known to be useful in modeling similar data, I would prefer to model this Illinois data using the weibull distribution, as it generally fit the data much better than the gamma. The only point of contention is the outlier, but we may be able to consider this point an anomaly in the dataset. Knowing this, we get the following mle estimates for the parameters:

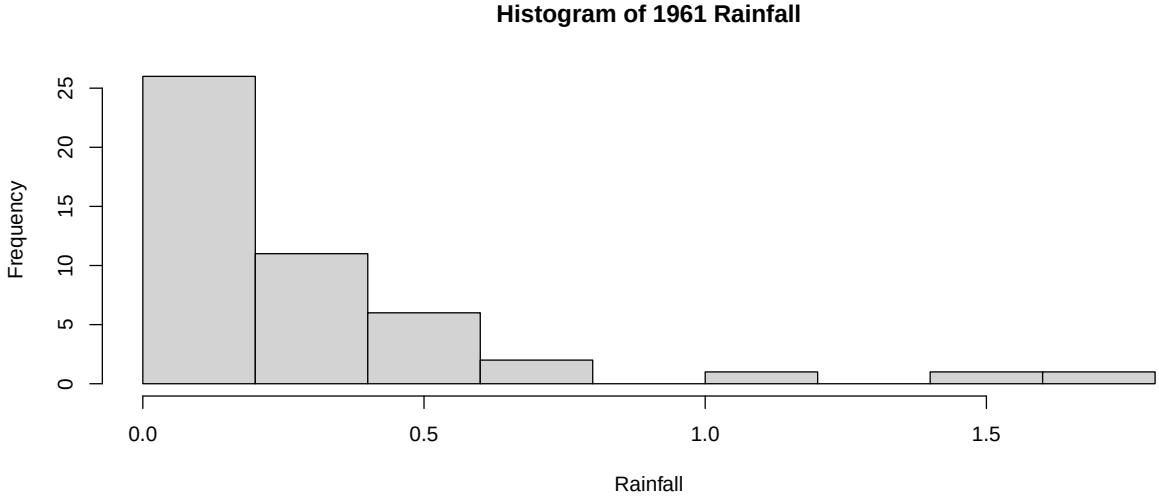
```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.5690986 0.02959323
## scale 0.1394868 0.01718016
## Loglikelihood:  189.4436   AIC:  -374.8873   BIC:  -368.0374
```

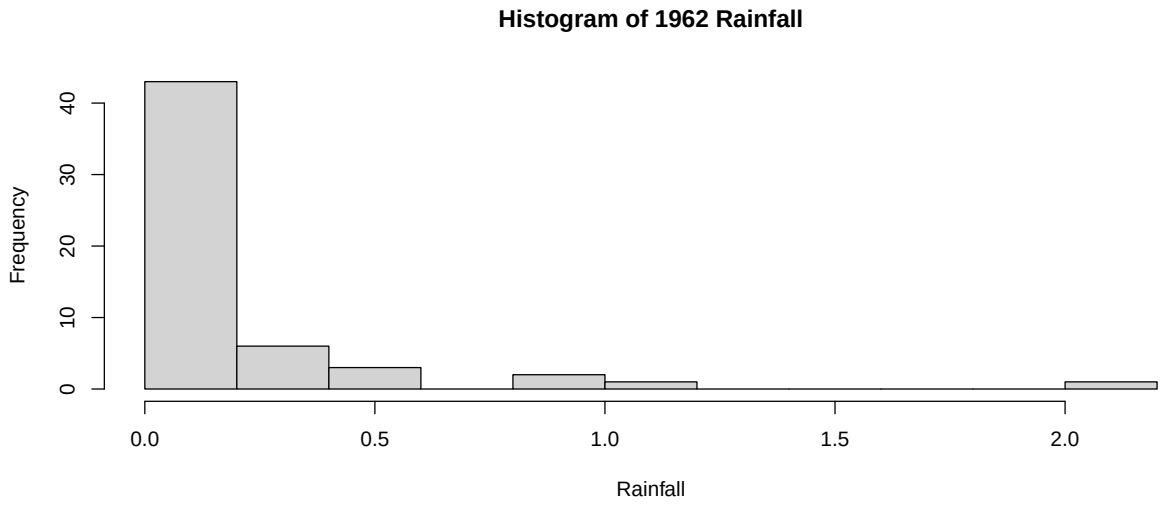
```
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3219421
## scale 0.3219421 1.0000000
```

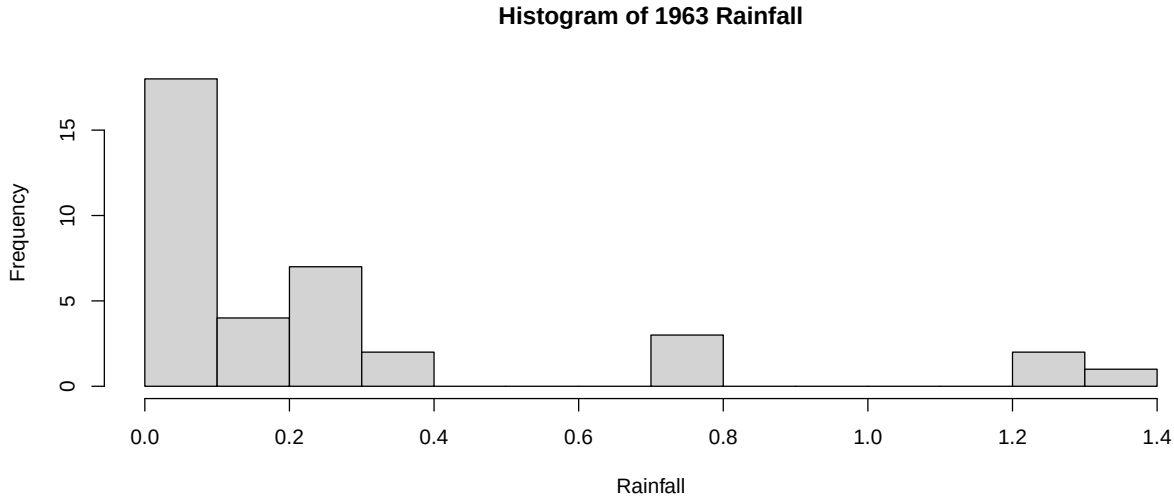
The shape parameter determines the shape of the distribution, while the scale parameter determines how spread out it is. The shape parameter of 0.57 being less than 1 gives the distribution its exponential like look, while the scale value of 0.14 being small keeps the distribution relatively compact. with a high right skew. Overall, I feel quite confident in this distribution and the parameter estimates as the empirical data seems to match it well from the above plots, and the estimates match the wanted shape of the distribution. The errors of the shape and scale parameters being relatively low at 0.03 and 0.02, respectively, is also quite encouraging.

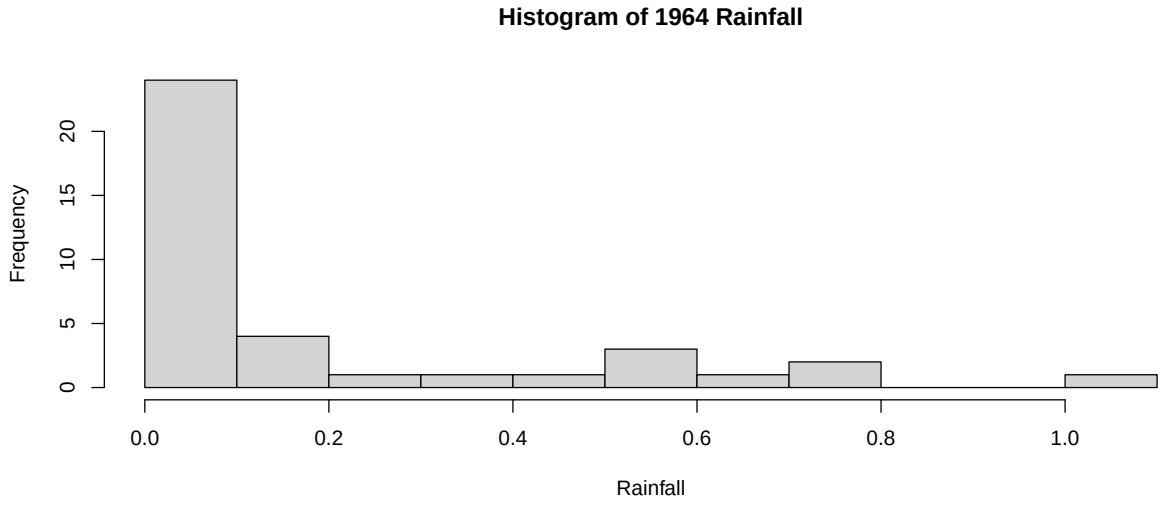
Using this distribution, we can attempt to locate wet years and dry years.





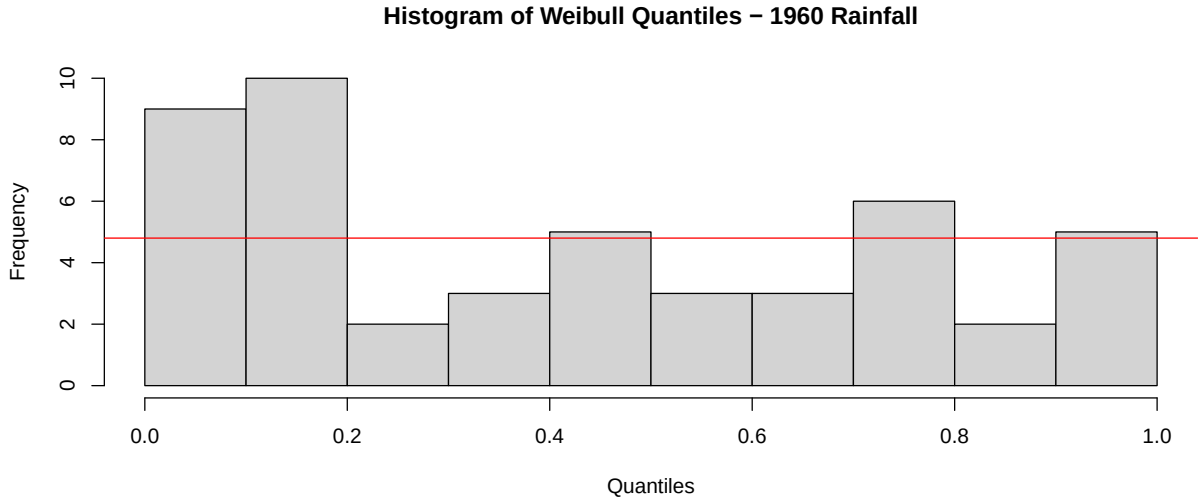

Histogram of 1962 Rainfall
Frequency
0
10
20
30
40
0.0
0.5
1.0
1.5
2.0
Rainfall

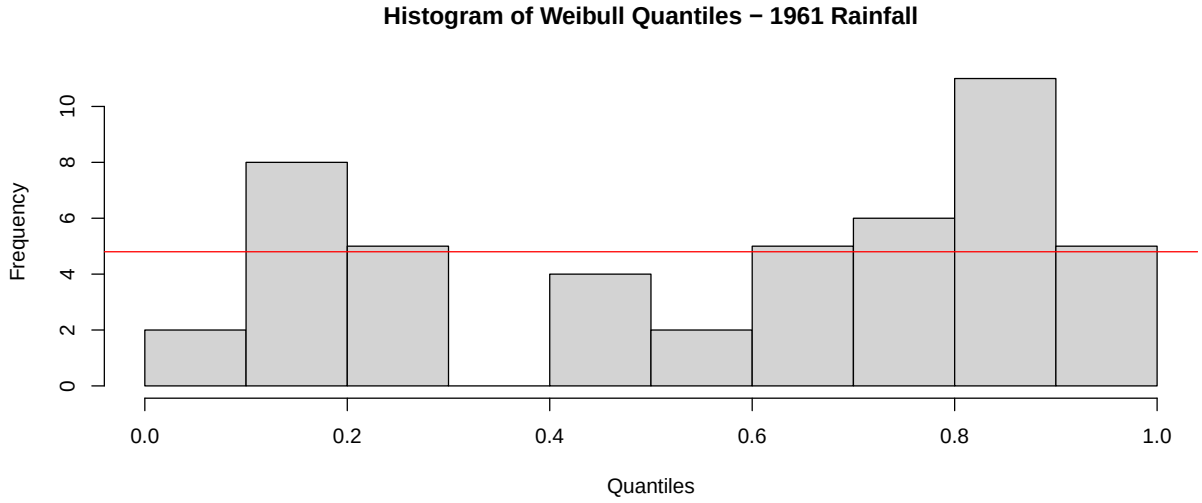

Histogram of 1963 Rainfall
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Rainfall

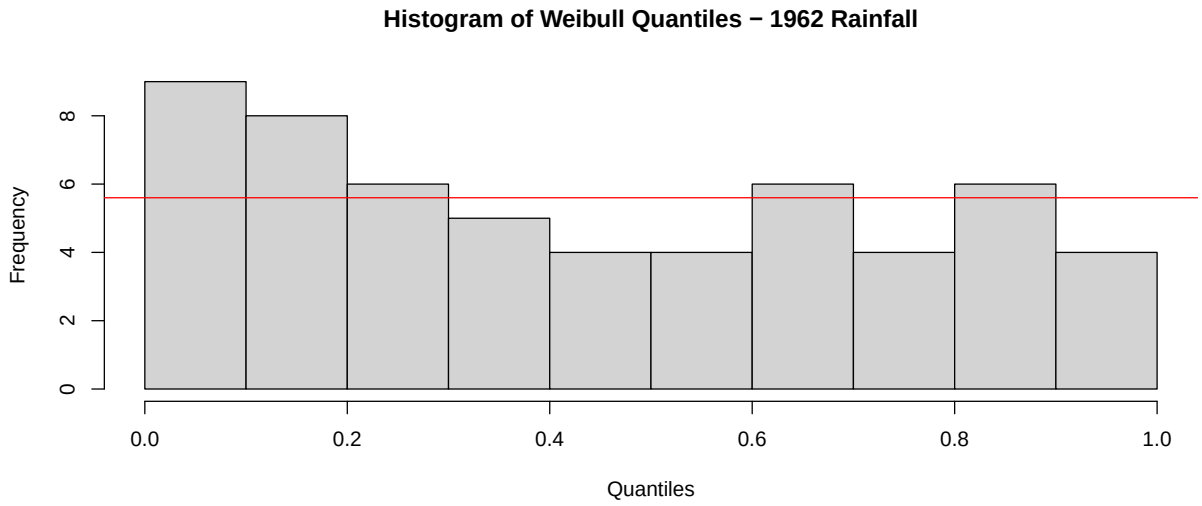

Histogram of 1964 Rainfall
Frequency
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
Rainfall

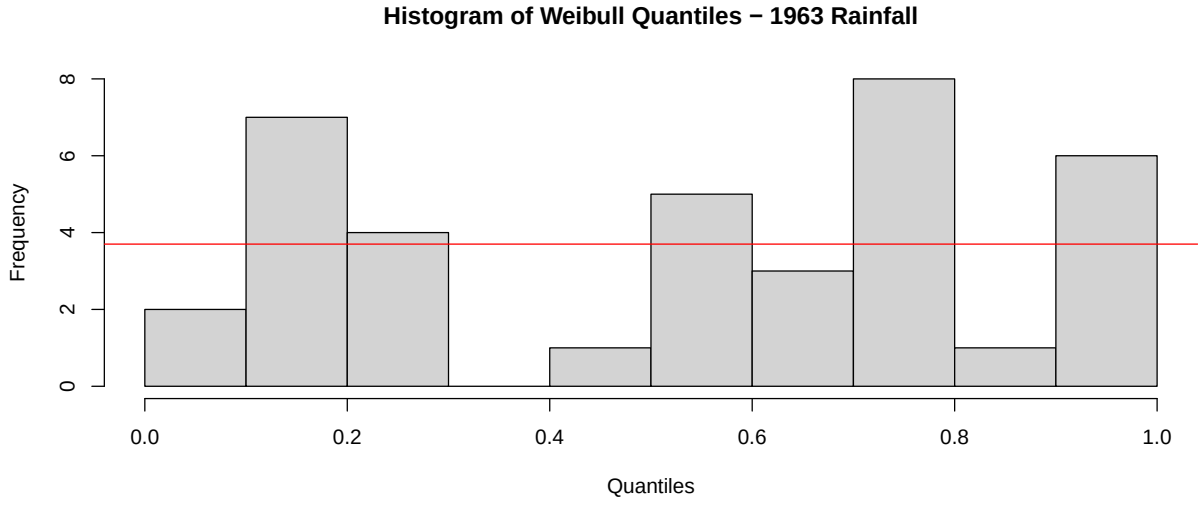
```
## Year 1960, Total Rain: 10.574 , Max Rainfall: 2.13 , Number of rainy days: 48

## Year 1961, Total Rain: 13.197 , Max Rainfall: 1.78 , Number of rainy days: 48

## Year 1962, Total Rain: 10.346 , Max Rainfall: 2.1 , Number of rainy days: 56

## Year 1963, Total Rain: 9.71 , Max Rainfall: 1.35 , Number of rainy days: 37

## Year 1964, Total Rain: 7.11 , Max Rainfall: 1.04 , Number of rainy days: 38
```

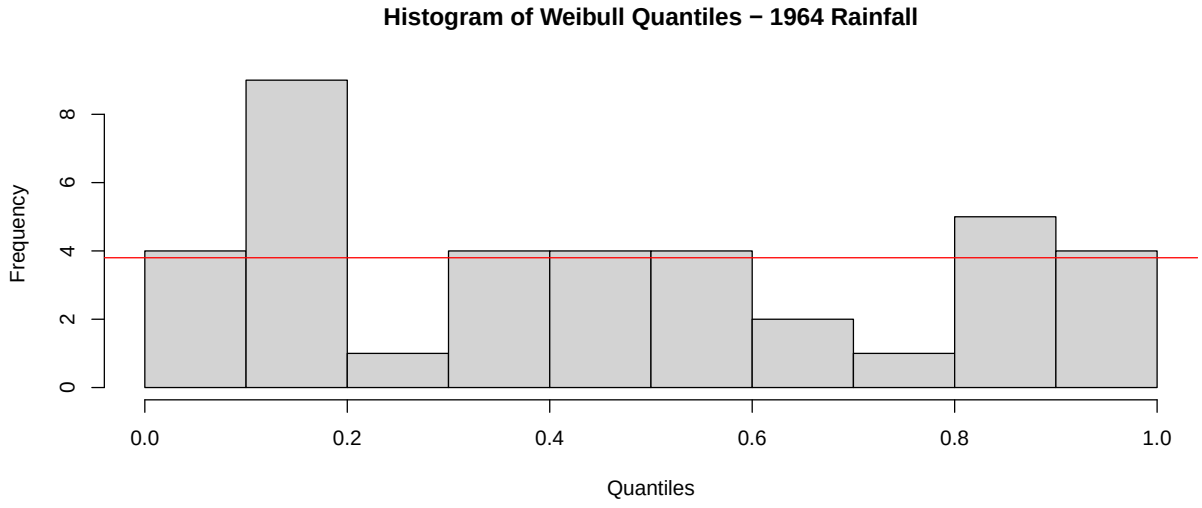
Based off the summary statistics alone, it would seem that both the number of rainy days and the amount of rain in each storm contributes to the wet/dry distinction.






Histogram of Weibull Quantiles − 1962 Rainfall
Frequency
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Quantiles


Histogram of Weibull Quantiles − 1963 Rainfall
Frequency
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Quantiles


Histogram of Weibull Quantiles − 1964 Rainfall
Frequency
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Quantiles

Here, the quantile values of each rain measurement based on the weibull(0.57,0.14) distribution are plotted as a histogram. This will allow us to see which years had more or less low-rain storms and high-rain storms. A year that was perfectly average would follow the distribution exactly, and would thus have each bar be approximately the same height. This theoretical average is represented by the red bars.

In the above histograms, we see that 1960 and 1964 had more low-rain storms than expected, while 1961 had more high-rain storms than expected. 1962 generally stayed at expected values, with a slight lean for low-rain storms. 1963 was rather volatile, with a mix of low-rain and high-rain storms, with fewer mid-rain storms; in general though, 1963 trended towards having more high-rain storms,

This result makes predicting wet and dry years off of the amount of rain per storm in a given year a bit dubious. While the high amount of high-rain storms signaled a high amount of total rain in 1961, the high amount of low-rain storms in 1960 still gave a relatively high total rainfall for the year. Similarly, 1964 having more low-rain days led to the total yearly rainfall being low, while the high amount of high-rain days for 1963 still led to low low total yearly rainfall.

Thus, it would appear that amount of rain per storm is important to determining wet and dry years, the amount of storms is equally if not more important. For instance, 1960 had the second most amount of rainy days, meaning that all of the low-rain storms added up to high amount of total rain on the year. In a similar manner, despite 1963 having a good amount of high-rain storms, it had the least amount of rainy days over the 5 years, dragging the total yearly rainfall down.

Overall, while these results are interesting, their generalization must be taken with caution. The data from the study is still rather small in size, consisting of only 5 consecutive years in an isolated region. Importantly, this data is also from about 60 years, with the state of the globe's climate changing dramatically within that time frame. However, the use of the weibull distribution (or even the gamma distribution) may still be valid to this day, as rain measurement systems still remain mostly the same, and the distribution of rainfall may still follow a similar trend.

A good next step for this analysis may be to repeat the study on some more current rain data. New data could be sampled, or it could be extracted from previous years from local weather stations. By doing this, we could see if the weibull distribution and the parameters still suffice, or if they could be updated for today's standards. More areas' rainfalls could also be studied, to see if the distribution is generalizable or is specific to certain climates. One would think regions more tropical than Illinois may have differing rainfall patterns, but evidence of the contrary would certainly be interesting.

For this final project overall, I learned more about estimating the distributions of functions based off empirical evidence. I also was able to practice evaluating models on different scales, and looking at the big picture in order to make decisions and analyses. In the future, I'll learn about more distributions, so as to better recognize them off empirical data in the future. I am also now more curious about estimation of parameters, and how the changes in their values may alter the distribution, and thus how we understand the data. To help with this, I may take on some projects in my own time, or at least be more aware of the subject in future work. I would also like to spend more time in the future understanding likelihood. Something I would probably do differently is spend some more time researching the topic before jumping into analysis.