MA677 Final Project - In All Likelihood

Yuyang Li

5/11/2022

Introduction

I chose task In All Likelihood.

4.25

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

integrand <- function(x,r,n) {
  x * (1 - F(x))^(r-1) * F(x)^(n-r) * f(x)
}
# In a sample of size n the expected value of
# the rth largest order statistic is given by:
E <- function(r,n) {
  (1/beta(r,n-r+1)) * integrate(integrand,-Inf,Inf, r, n)$value
}
# function of approximation
approxi = function(i,n){
  m = (i-1/3)/(n+1/3)
  return(m)
}
```

Then we test the value of approximation and order statistic respectively.

```
# n = 5
E(2.5,5)
```

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

```
approxi(2.5,5)
```

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

```r
# n = 10
E(5,10)
```

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

```r
approxi(5,10)
```

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

The values of median approximation and order statistics are similar.

4.27

```r
jan1940 = 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)

jul1940 = 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)

```r
summary(jan1940)
```

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

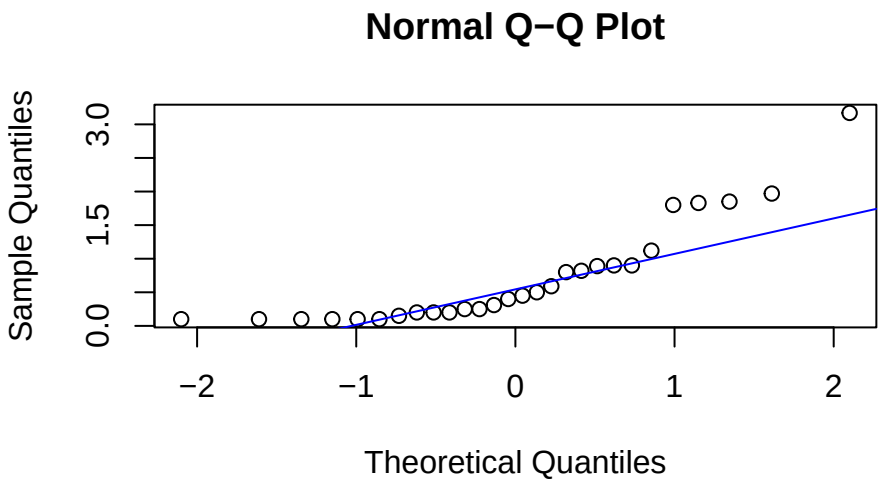
```r
summary(jul1940)
```

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

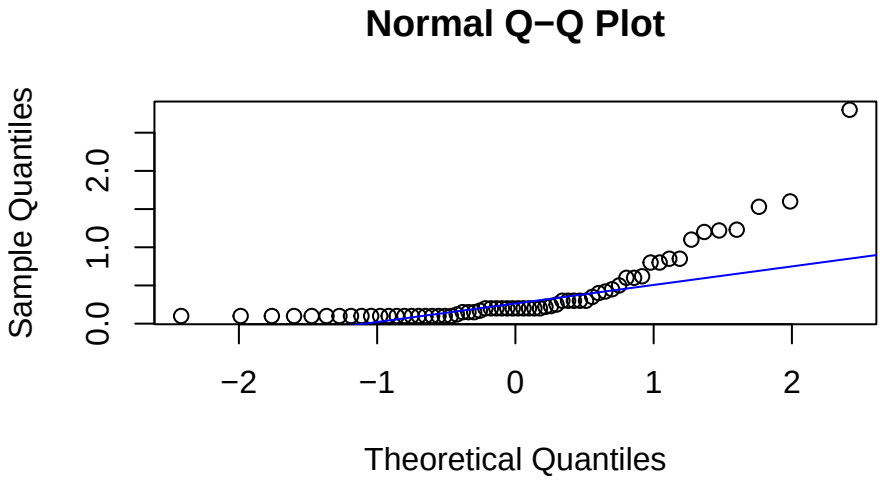
The overall statistical value in January 1940 is higher than in July 1940.

(b)

```r
# Get QQ-plots for two data sets
qqnorm(jan1940)
qqline(jan1940,col = "blue")
```




```r
qqnorm(jul1940)
qqline(jul1940,col = "blue")
```



From QQ-plot I found that compared with normal distribution we cannot get a value smaller than the theoretical small value, but the largest value would larger than be expected, which means a long tail. Then I consider it as a skewed distribution and the Gamma distribution might be reasonable.

(c)

```
# fit gamma model to two months
set.seed(2022)
fit.gamma_jan = fitdist(jan1940,distr = "gamma",method = "mle")
```

```
fit.gamma_jul = fitdist(jul1940,distr = "gamma",method = "mle")
```

```
# Report the MLEs and standard errors
# Jan1940
fit.gamma_jan$estimate # MLEs
```

```
##    shape     rate
## 1.056222 1.467650
```

```
fit.gamma_jan$sd #standard errors
```

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

```
# Jul1940
fit.gamma_jul$estimate # MLEs
```

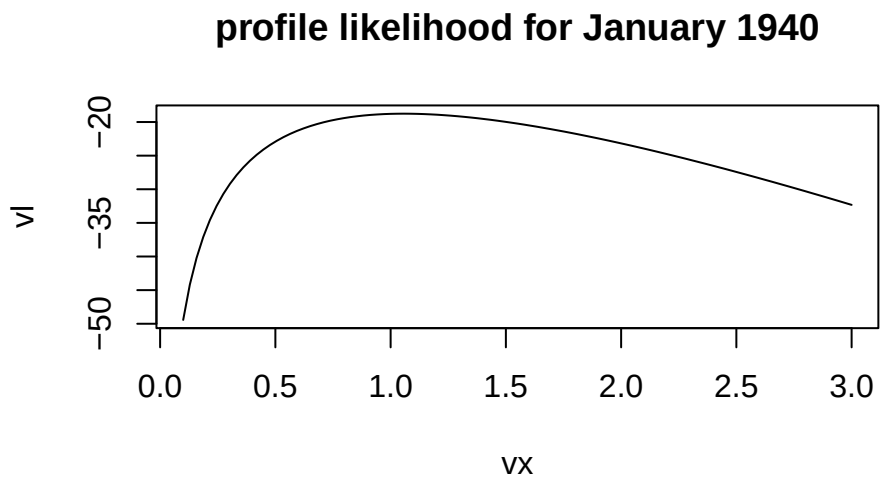
```
##    shape     rate
## 1.196419 3.043403
```

```
fit.gamma_jul$sd #standard errors
```

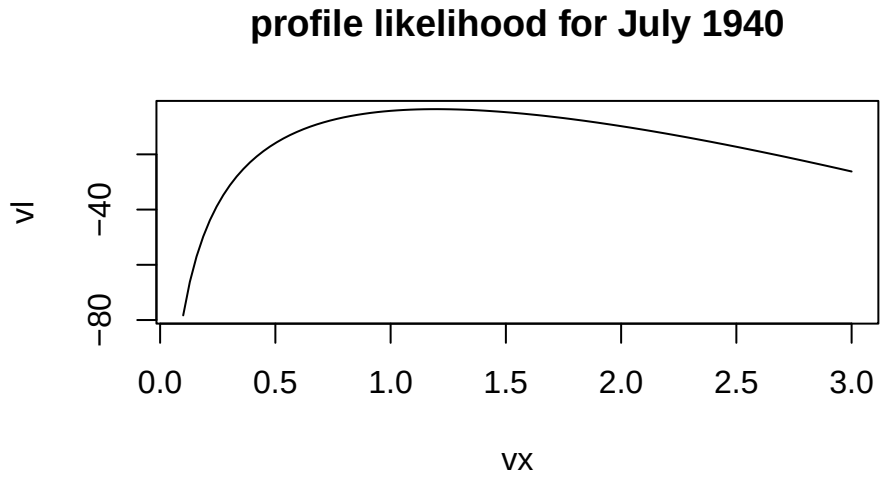
```
##     shape      rate
## 0.1891196 0.5936302
```

The estimated value of alpha and beta in January is lower than in July. For standard error, the January model has a smaller sd of beta but July's sd of alpha is smaller.

```
# profile likelihoods for January 1940
# reference:https://www.r-bloggers.com/2015/11/profile-likelihood/
x = jan1940
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=101)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main = "profile likelihood for January 1940")
```

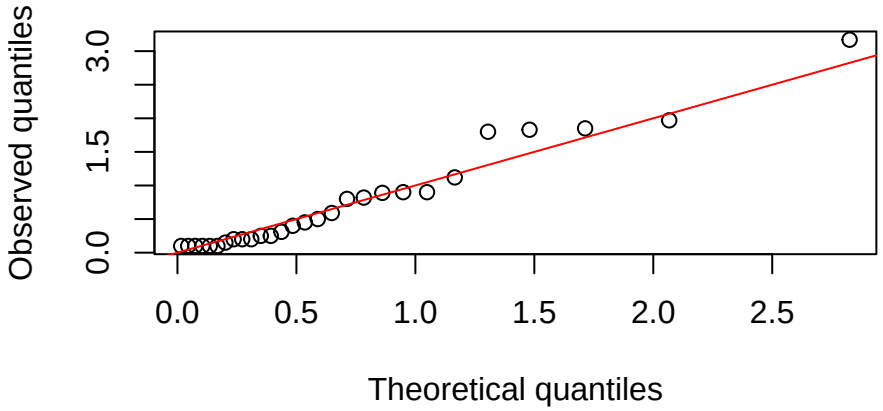



```
# profile likelihoods for July 1940
x = jul1940
prof_log_lik=function(a){
  b=(optim(1,function(z) -sum(log(dgamma(x,a,z)))))$par
  return(-sum(log(dgamma(x,a,b))))
}
vx=seq(.1,3,length=101)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main = "profile likelihood for July 1940")
```

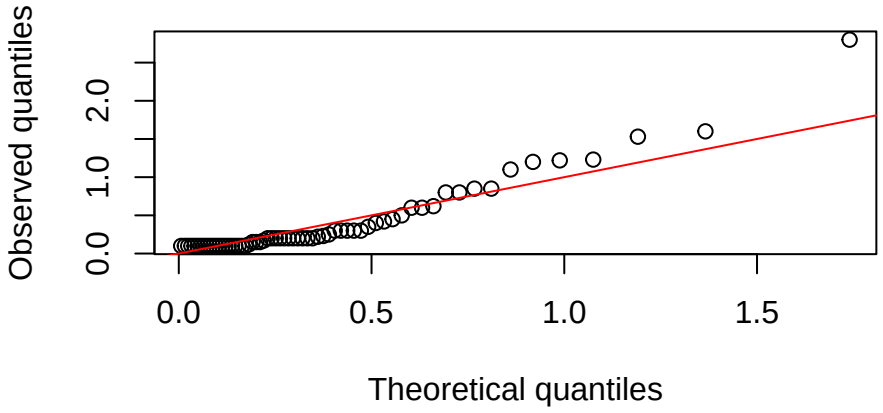


(d)

```
set.seed(2017);
# Jan1940 are Gamma distributed data
x <- jan1940
# Sort x values
x <- sort(x)
# Theoretical distribution
x0 <- qgamma(ppoints(length(x)), shape = fit.gamma_jan$estimate[1],
             rate = fit.gamma_jan$estimate[2]);
plot(x = x0, y = x, xlab = "Theoretical quantiles", ylab = "Observed quantiles");
abline(a = 0, b = 1, col = "red");
```



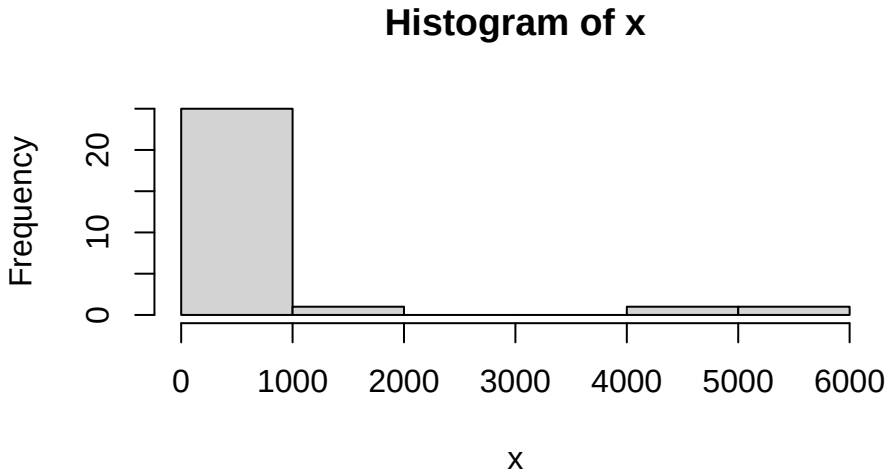

```
set.seed(2017);
# Jul1940 are Gamma distributed data
x <- jul1940
# Sort x values
x <- sort(x)
# Theoretical distribution
x0 <- qgamma(ppoints(length(x)), shape = fit.gamma_jul$estimate[1],
             rate = fit.gamma_jul$estimate[2]);
plot(x = x0, y = x, xlab = "Theoretical quantiles", ylab = "Observed quantiles");
abline(a = 0, b = 1, col = "red");
```



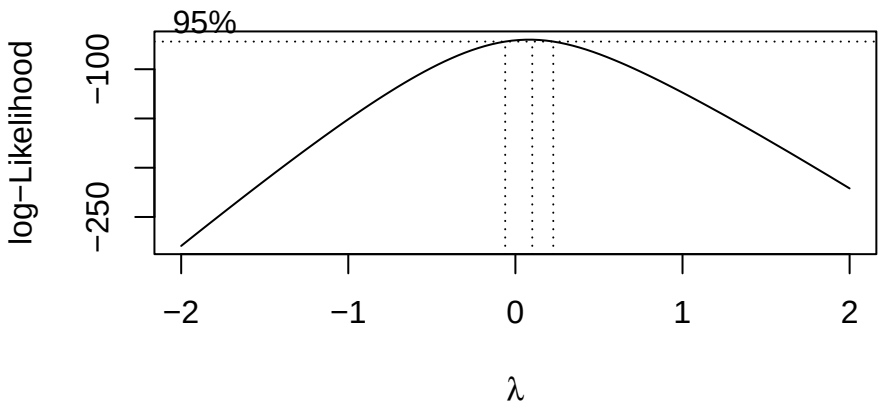
Through the gamma QQ-plot it seems the gamma model is appropriate.

4.39

```
x = 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(x)
```

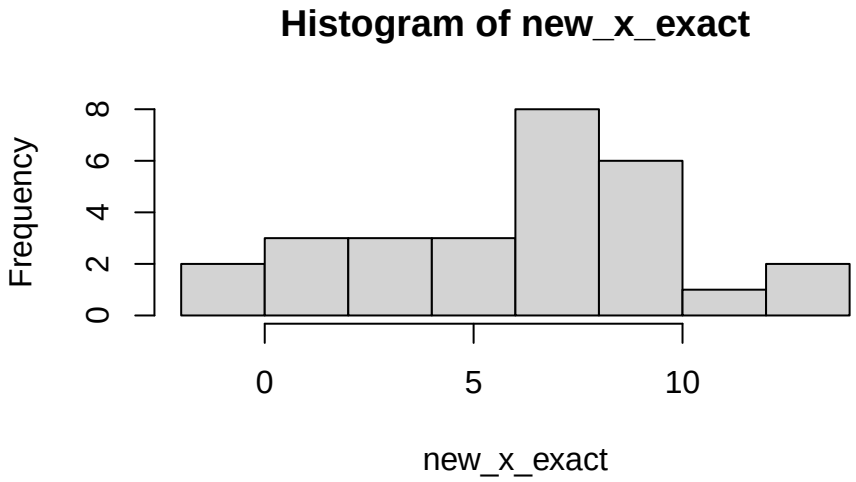



```
# Apply box-cox transform
library(MASS)
b = boxcox(lm(x~1))
```

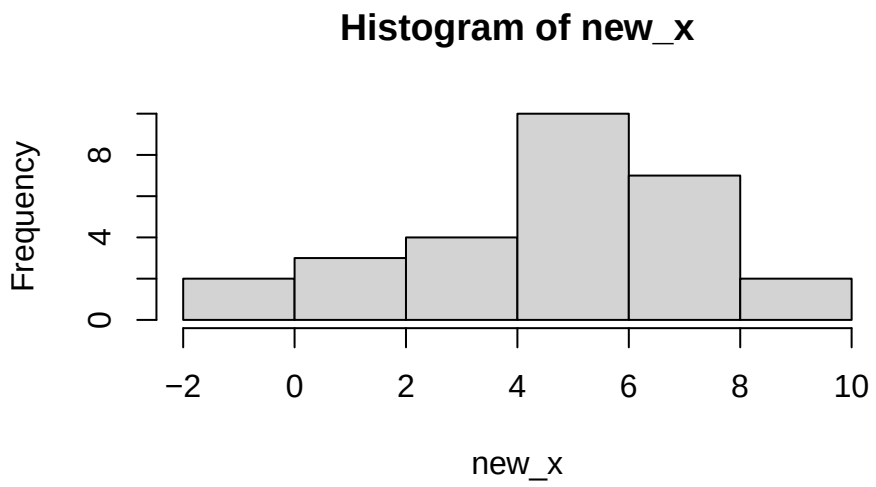



```
# Exact lambda
lambda <- b$x[which.max(b$y)] # 0.1010101
```

```
# Transform x to new x
new_x_exact <- (x ^ lambda - 1) / lambda
hist(new_x_exact)
```




```
# If we choose lambda = 0, apply the logarithmic transformation
new_x = log(x)
hist(new_x)
```



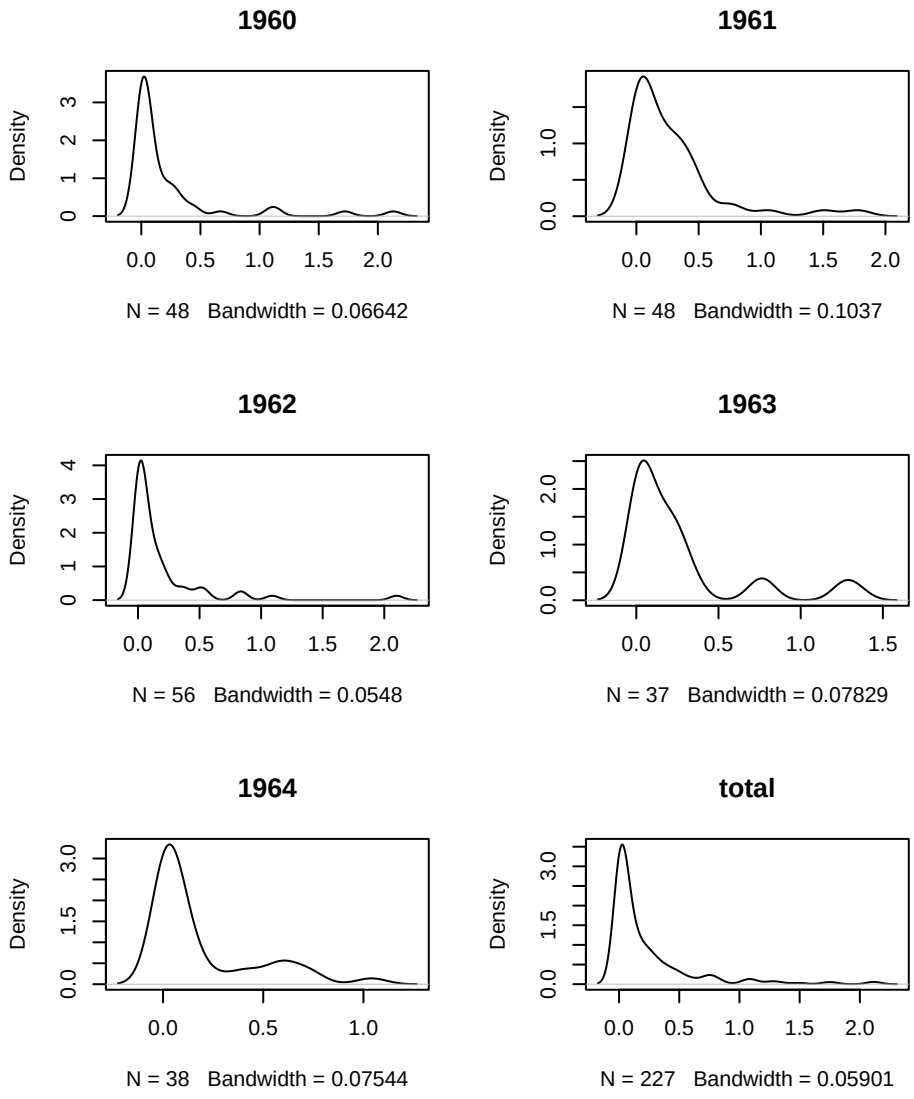
After both transformations, the data are presented more like a normal distribution.

Rainfall analysis

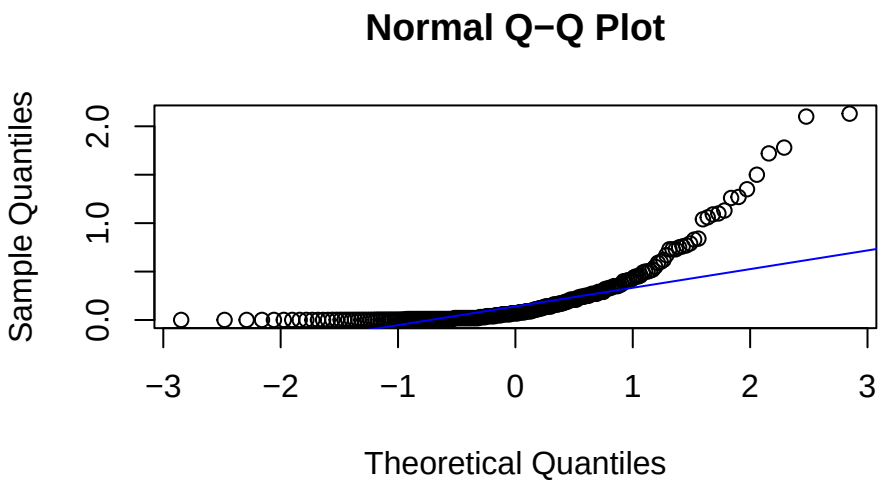
Part 1

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("Illinois_rain_1960-1964(1).xlsx")
# Get density plots for each year and total data
par(mfrow = c(3,2))
plot(density(rain$`1960`%>%na.omit()),main = "1960")
plot(density(rain$`1961`%>%na.omit()),main = "1961")
plot(density(rain$`1962`%>%na.omit()),main = "1962")
plot(density(rain$`1963`%>%na.omit()),main = "1963")
plot(density(rain$`1964`%>%na.omit()),main = "1964")
total_rain = unlist(rain)%>%na.omit()
plot(density(total_rain),main = "total")
```




```
# Get QQ-plot for total data
qqnorm(total_rain)
qqline(total_rain,col = "blue")
```



Through the density plot and QQ-plot for data I found it seems like gamma distribution same as 4.27. Then I use total data to apply fitdist fitting gamma model with MLE and MSE.

```
total_rain = as.numeric(total_rain)
fit_mle = fitdist(total_rain,distr = "gamma",method = "mle")
fit_mse = fitdist(total_rain,distr = "gamma",method = "mse")
```

```
summary(fit_mle)
```

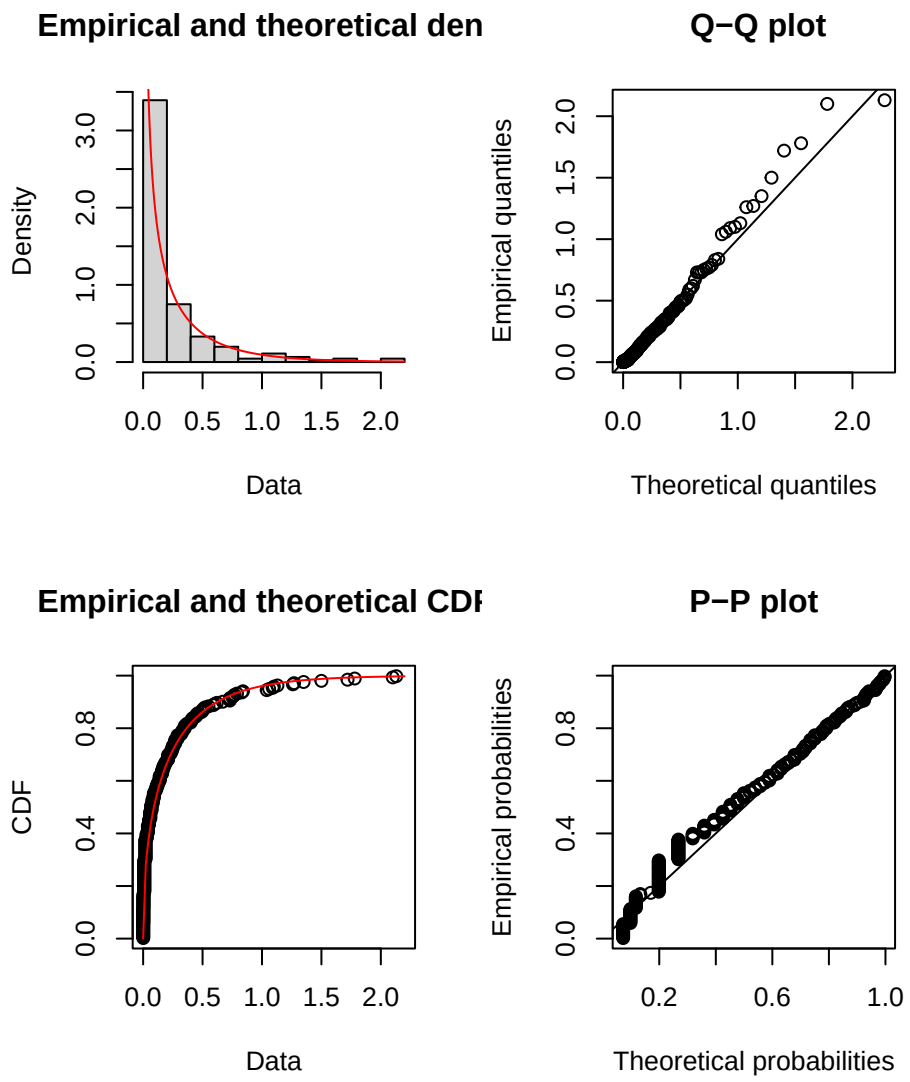
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 0.4408386  0.0337663
## rate  1.9648409  0.2474440
## Loglikelihood:  185.3477   AIC:  -366.6954   BIC:  -359.8455
## Correlation matrix:
##           shape      rate
## shape 1.0000000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

```
# Use bootdist to get confidence interval of parameters
summary(bootdist(fit_mle))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4427032 0.3799036 0.5199596
## rate  1.9603479 1.5198481 2.5459979
```

From the result of the model, the MLE of alpha is 0.4408386 with a standard error of 0.0337663, MLE of beta is 0.0337663 with a standard error of 0.0337663. Additionally, the 95% confidence interval of alpha is [0.0337663, 0.5225162] and for beta is [1.5496223, 2.5292835]

```
plot(fit_mle)
```



Through these figures, I believe that using gamma distribution and MLE to fit the model is appropriate.

Part 2

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?

Since I identified data as gamma distribution, so the mean of total data was supposed to be $\mu = \frac{\alpha}{\beta}$

```
# Average rainfall for five years
mean_total = fit_mle$estimate[1]/fit_mle$estimate[2]
# Average storms number for five years
num_total_avg = length(total_rain)/5
```

```r
# Average rainfall of each year
mean = c(apply(rain,2,mean,na.rm=TRUE),mean_total)%>%round(4)
names(mean)[6] = "Total"
# Individual storm of each year
num = c(apply(!is.na(rain),2,sum,na.rm = TRUE),num_total_avg)
names(num)[6] = "Total"
knitr::kable(rbind(mean,num))
```

	1960	1961	1962	1963	1964	Total
mean	0.2203	0.2749	0.1848	0.2624	0.1871	0.2244
num	48.0000	48.0000	56.0000	37.0000	38.0000	45.4000

According to the average rainfall of each year vs the total five years, I defined 1960, 1962,1964 are dry years since their annually average rainfalls are lower than the total years. The years 1961, and 1963 are wet years conversely. For the wet year 1961, I think the reason for wet was more storms in this year because the number of storms was 48, larger than the average number of storms over the five years. However, In the year 1963, it might be caused by more rainfall per storm because the number of storms was only 37, much smaller than five years average.

Part 3

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? An article by Floyd Huff, one of the authors of the 1967 report is included.

I thought the generalization of my results was limited since the data set was small, and only contained five years' data. For the next step, I can collect more annual rainfall data to verify my conclusion and make my model more generalizable. Also, I could try data outside Illinois to figure out the rainfall difference among places. And maybe the appearance of the wet and dry year is regular, more data is necessary for me to conduct more analysis. In Huff's paper, he provided a discussion of various factors that might influence the rainfall like landform and position. He discussed the variable of rainfall over years as well.

Summary

In this project, I learned a lot about likelihood and kinds of methods to analyze data and develop models, identify distribution. Absolutely, for my next step, I will dig into this book and gain more statistical knowledge. Also, I found that practice is the best way to understand an intricate theory. Therefore, I would like to practice more about how to apply statistics to data using R.

I also encountered some problems like how to identify the distribution of data was gamma distribution. At first I considered histogram but it's hard to get the identification. To solve this issue, I searched numerous websites and finally found my result is using density plot.

Lastly, thanks for the help from Zening Ye, he taught me how to identify the wet or dry year.