MA677 Final Project

Chen Xu(U49903384)

May 10 2022

#Introduction:

I choose to do the problems based on In All Likelihood.

###Problem 4.25

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

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

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

# Approximation function
approxf <-function(i,n){
  m <- (i-1/3)/(n+1/3)
  return(m)
}
```

```
# n = 5
(median_U(2,5) + median_U(3,5))/2
```

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

```
(approxf(2,5) + approxf(3,5))/2
```

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

```
#n = 10
(median_U(5,10) + median_U(6,10))/2
```

```
## [1] 0.5
```

```
(approxf(5,10) + approxf(6,10))/2
```

```
## [1] 0.5
```

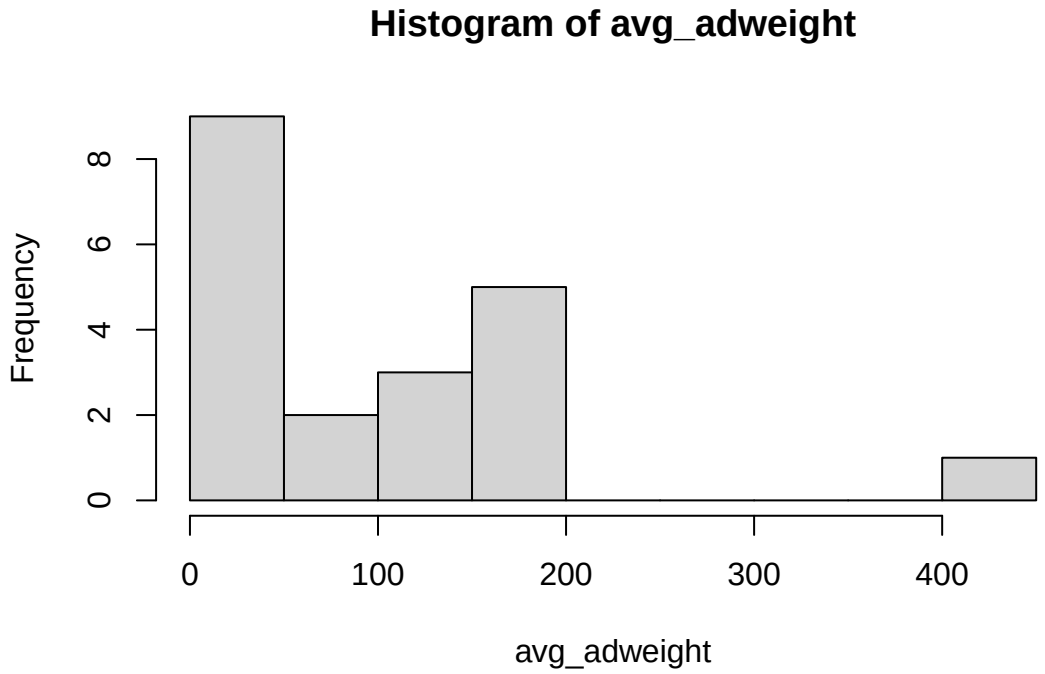
According to the results I got, I can say that:

$$\text{median}\left\{U_{(i)}\right\} \approx \frac{i-1/3}{n+1/3}$$

##Problem 4.39

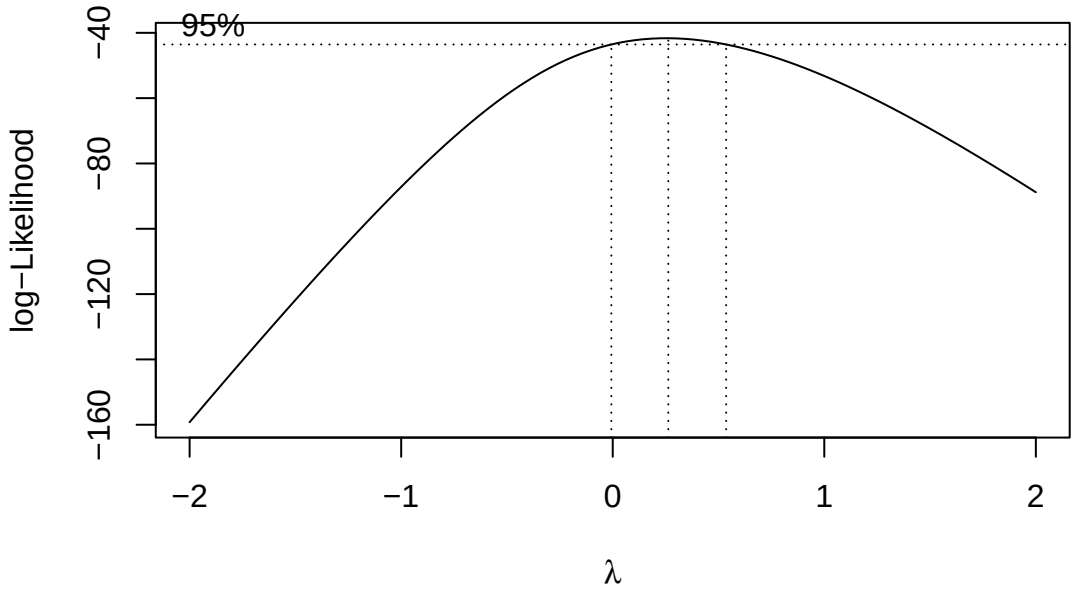
```
# Build the data set for  28 species of animals
avg_adweight <- c(0.4,1.0,1.9,3.0,5.5,8.1,12.1,25.6,50.0,56.0,70.0,115.0,115.0,119.5,154.5,157.0,175.0,

# See the distribution of the data set
hist(avg_adweight)
```



Conduct boxcox transformation

```
bc_transformation <- boxcox(lm(avg_adweight ~ 1))
```



As can see from the box-cox transformation plot, the 95 confidence interval for λ does not include 1, so a transformation is appropriate.

The next step, I extract the optimal λ and make transformation to the original data set.

Exact lambda

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

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

Check the distribution of transformed data set:

```
transformed_data <- (avg_adweight ^ lambda - 1) / lambda
hist(transformed_data)
```

##Problem 4.27

Create data sets:

```
Jan.1940 <-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)

Jul.1940 <-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)

```
summary(Jan.1940)
```

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

```
summary(Jul.1940)
```

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

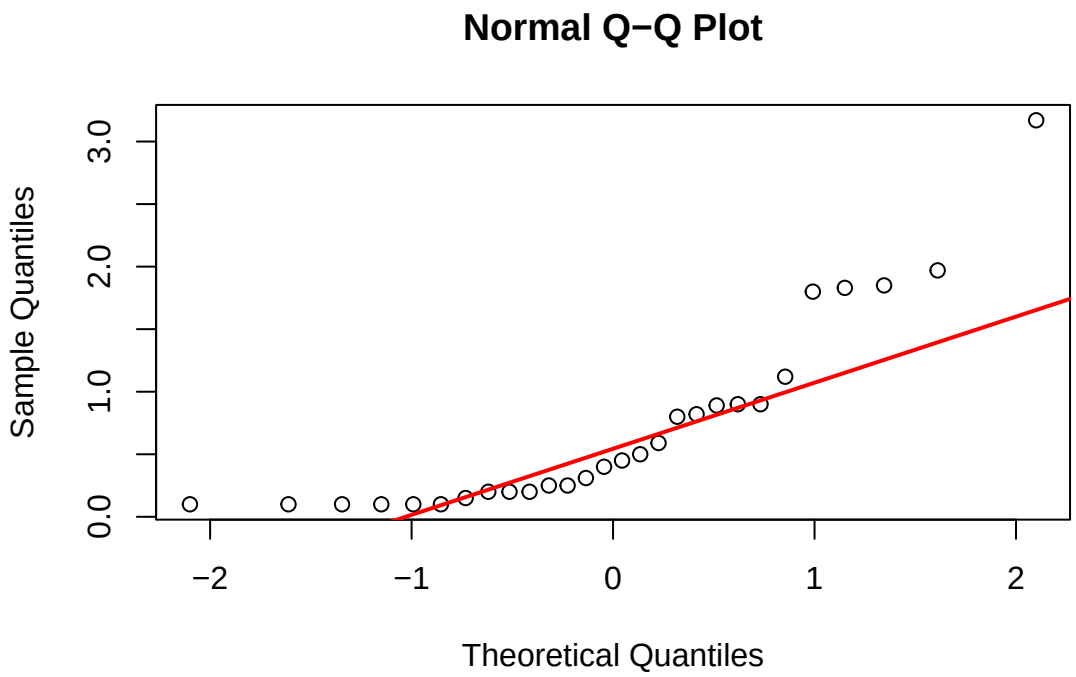
```
##  0.1000  0.1000  0.2000  0.3931  0.4275  2.8000
```

After comparing the summary statistics for the two months data about the average amount of rainfall (in mm/hour) per storm in a series of storms in Valencia, southwest Ireland. Expect the minimum value, every summary statistics of January 1940 are higher than July 1940.

###(b)

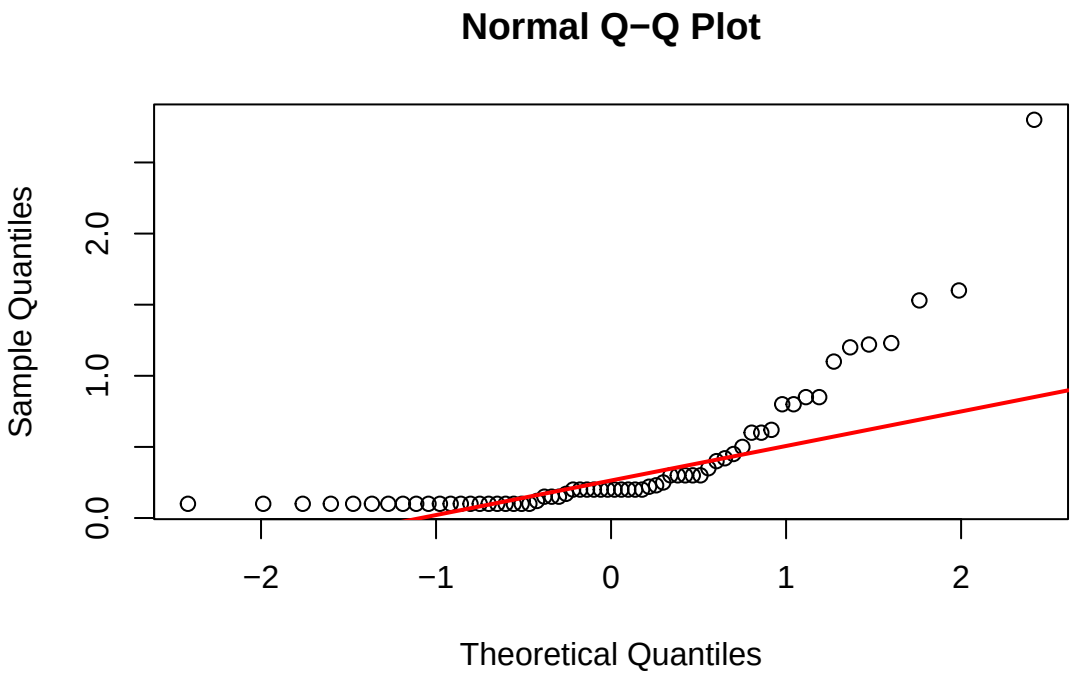
QQ-plot of January 1940:

```
qqnorm(Jan.1940, pch = 1)
qqline(Jan.1940, col = "red", lwd = 2)
```



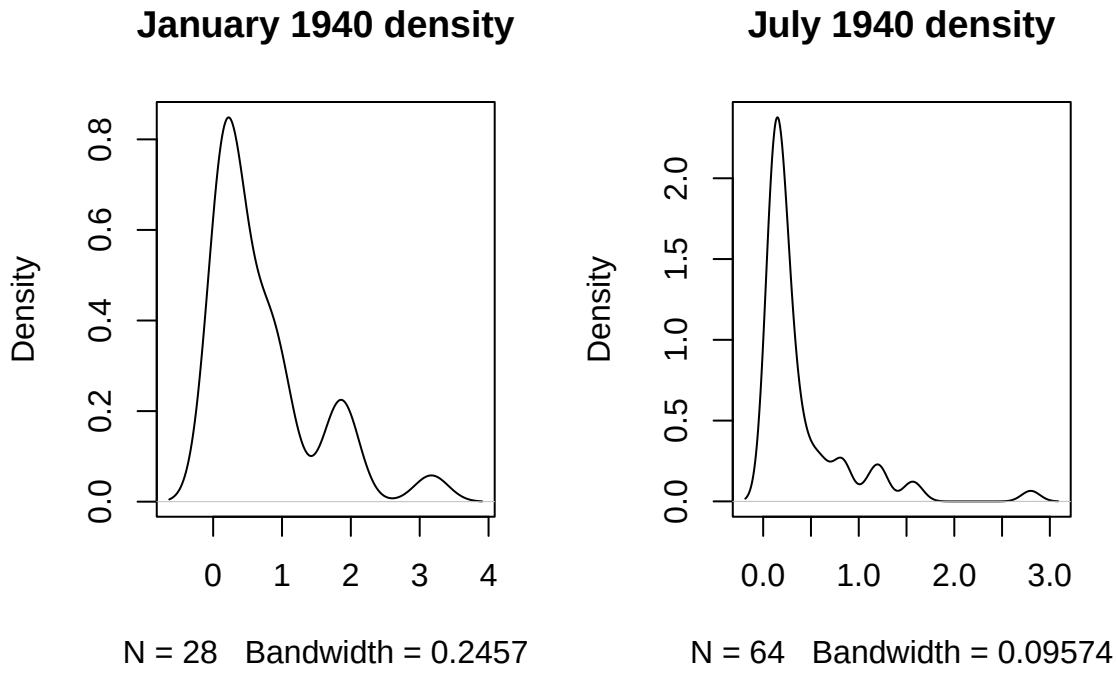
QQ-plot of July 1940:

```
qqnorm(Jul.1940, pch = 1)
qqline(Jul.1940, col = "red", lwd = 2)
```



Density plot(shape):

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



The QQ-plots show that the sample doesn't follow normal distribution. From the density shape plot, these data looks like gamma distribution. Therefore, the model I suggest is gamma distribution.

###(c)

```r
Jan.fit <- fitdist(Jan.1940,'gamma','mle')
summary(Jan.fit)
```

```
## 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
July.fit <- fitdist(Jul.1940,'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
```

For MLE, exponentiate loglikelihood into MLE:

```r
exp(Jan.fit$loglik)
```

```
## [1] 7.11117e-09
```

```r
exp(July.fit$loglik)
```

```
## [1] 0.02638693
```

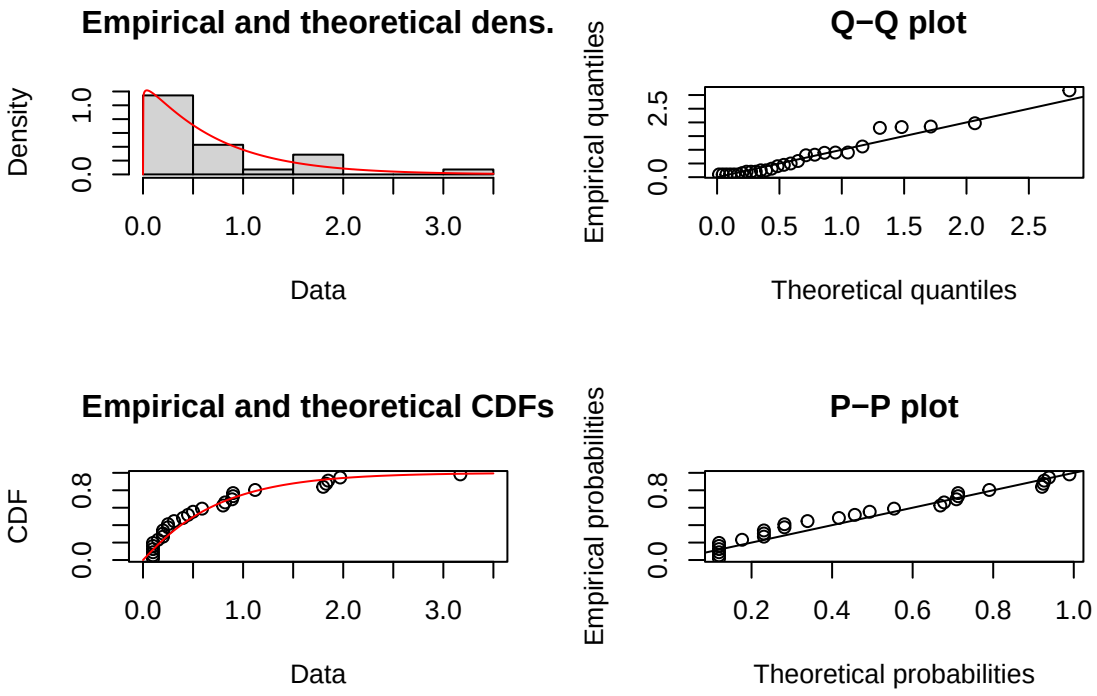
As we can see from the value of MLE. The MLE from data of July 1940 is higher than the one of January 1940. And from the AIC and BIC, we can know that the model for data of July 1940 is better than the model for data of June 1940.

Parameter comparison: The model for data of July 1940 has smaller alpha. The model for data of January has larger beta.

###(d)

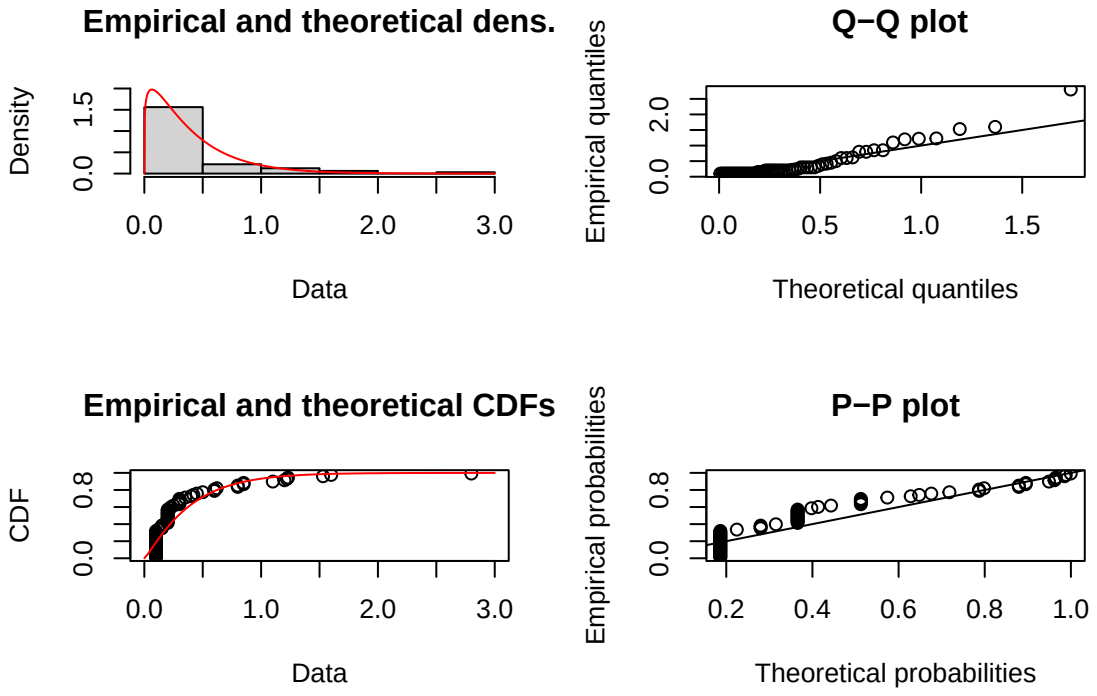
Check the adequacy of the gamma model for January 1940 using a gamma QQ-plot.

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



Check the adequacy of the gamma model for July 1940 using a gamma QQ-plot.

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



After checking the adequacy of two gamma model, I may conclude that the model for data of July 1940 is better.

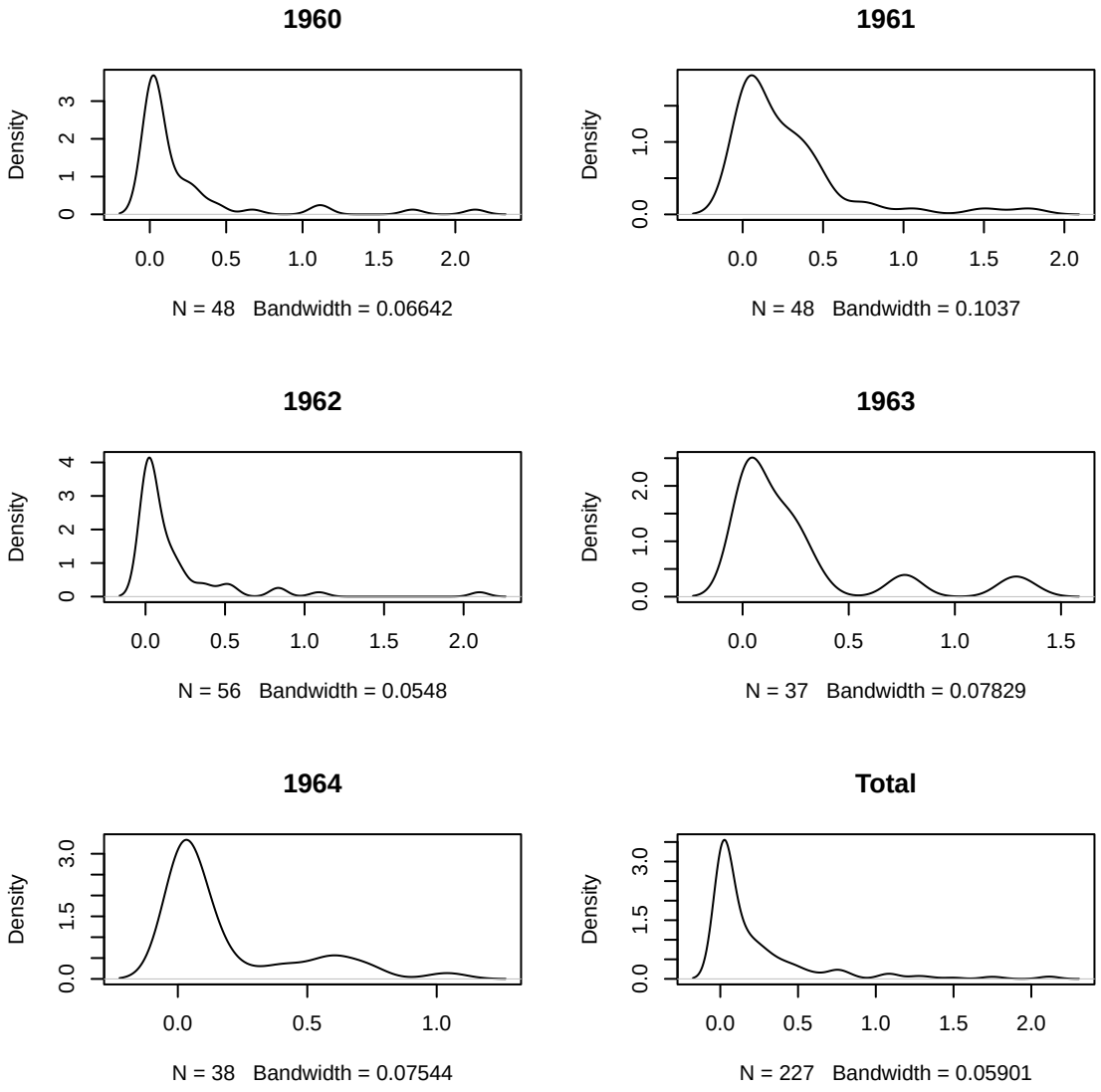
Illinois rain

###(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.

First step, check the density shape for the rainfall from southern Illinois 1960-1964 and the total 5 years to determine what distribution model to use.

```
il_rain <- read.xlsx('Illinois_rain.xlsx')
par(mfrow = c(3, 2))
density(il_rain$`1960` %>% na.omit()) %>% plot(main='1960')
density(il_rain$`1961` %>% na.omit()) %>% plot(main='1961')
density(il_rain$`1962` %>% na.omit()) %>% plot(main='1962')
density(il_rain$`1963` %>% na.omit()) %>% plot(main='1963')
density(il_rain$`1964` %>% na.omit()) %>% plot(main='1964')
density(unlist(il_rain) %>%  na.omit()) %>% plot(main='Total')
```

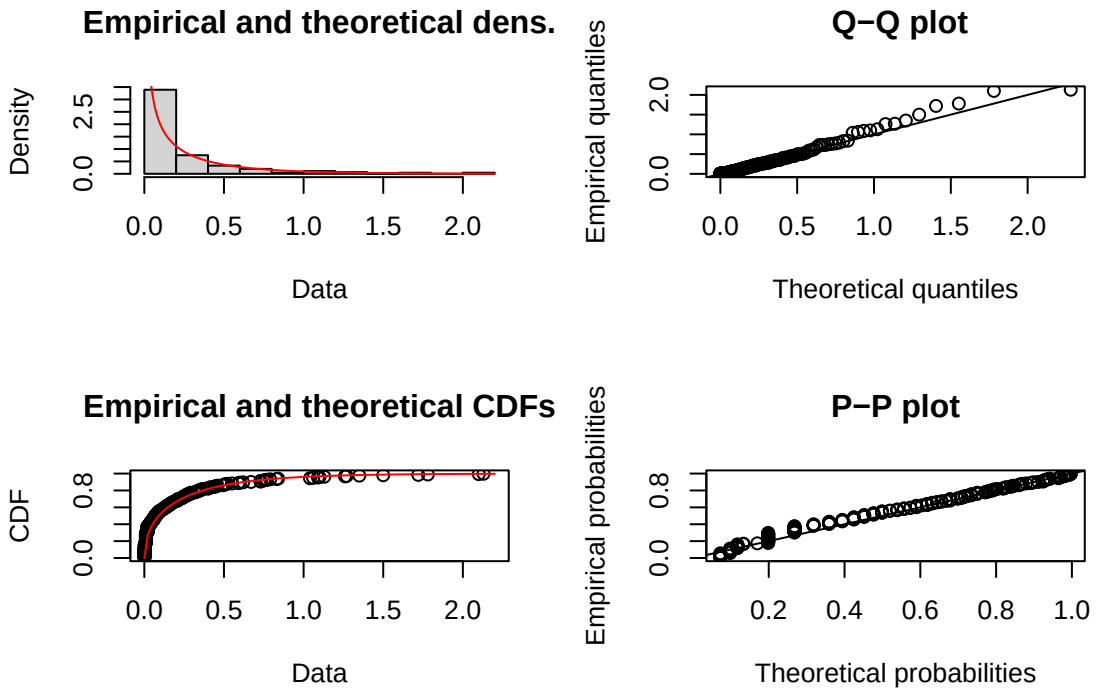


From the distribution density shape, we can see gamma distribution should be the better choice. Then I use the whole data set to fit a gamma model.

```
fit1 <- fitdist(unlist(il_rain) %>%  na.omit() %>% c(),'gamma',method='mle') #MLE estimation
```

Use gamma QQ-plot to check if gamma distribution a good fit for the rainfall data:

```
plot(fit1)
```



According to the Empirical the theoretical CDFs and P-P plot, the data points almost excellent fit the gamma distribution. Although in Q-Q plot, the points after 1.0 theoretical quarantines are a little bit discrete from the gamma distribution. Generally, the gamma distribution was a good fit for the total rainfall data.

In order to show how confidence I am about the identification of the distribution and the accuracy of my parameter estimates, I will compare the 95% confidence interval between the distribution using MLE and MSE using bootstrap.

```
set.seed(1234)
# calculate MSE and MLE
fit2 <- fitdist(unlist(il_rain) %>%  na.omit() %>% c(),'gamma',method='mse') #MLE estimation
boot_mse  <-  bootdist(fit2)
summary(boot_mse)
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.7154716 0.6279317 0.8527069
## rate  1.3378817 1.0810701 1.6691432
```

```
boot_mle  <-  bootdist(fit1)
summary(boot_mle)
```

```
## Parametric bootstrap medians and 95% percentile CI
##         Median      2.5%     97.5%
## shape 0.442677 0.3843748 0.5178236
## rate  1.986727 1.5576238 2.5654712
```

For MSE, the 95% confidence interval of shape from bootstrap sample is (0.6279317,0.8527069), the rate is (1.0810701,1.6691432).

For MLE, the 95% confidence interval of shape from bootstrap sample is (0.3843748,0.5178236),the rate is

(1.5576238, 2.5654712).

Apparently, the MLE estimates have narrow 95% and thus lower variances. I would suggest MLE being the estimator because it has lower variance.

The confidence interval indicates that the estimation is reliable.

###(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?

In order to compare wet years and dry years, I first calculate the average storm rainfall through the five year as a baseline, and calculate each year's average storm rainfall.

```
rf_mean <- fit1$estimate[1]/fit1$estimate[2]
rf_yealy <- apply(il_rain,2,mean,na.rm =TRUE)
avg_storm_rf <- c(rf_yealy,rf_mean %>% as.numeric() %>% round(4))
names(avg_storm_rf)[6]= '5 year avg'
avg_storm_rf
```

```
##       1960       1961       1962       1963       1964 5 year avg
##  0.2202917  0.2749375  0.1847500  0.2624324  0.1871053  0.2244000
```

Then generating the number of storms for each year:

```
num_storm<-c(nrow(il_rain)-apply(is.na(il_rain),2,sum),'/')
num_storm
```

```
## 1960 1961 1962 1963 1964
## "48" "48" "56" "37" "38"  "/"
```

Use table to show the result:

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

Compared to average storm rainfall each year with the baseline average, the year 1962, 1964 were dry years, the year 1961 and 1963 were wet years. The year 1960 is a normal year. But when we compare the number of storms for each, we can conclude that the year 1962 was a wet year, the year 1963 and 1964 were dry years and the year 1960 and 1961 were normal years. As a result, we may conclude that more storms don't necessarily result in wet year and more average rainfall in individual storm don't necessarily result in wet year, both of these reasons have influence on whether a year is a wet or dry year.

###(3)

To what extent do you believe the results of your analysis are generalizable?

From my perspective, 5-year data is not enough for my analysis to be generalizable. From Floyd Huff's Time Distribution Rainall in Heavy Storms in 1967. He used Network data for the 11-year period 1955 - 1966.

What do you think the next steps would be after the analysis?

From my perspective, the possible next steps are: 1. Build a reusable database to collect and track more years of storm rainfall data. 2. Fit the gamma distribution with more data. 3. May build a predictive model to validate our previous descriptive analysis.

##Citation:

1. I consulted this final project from Yuli Jin. From Yuli Jin, I learnt to use bootstrap to generate confidence interval for gamma distribution
2. Order statistics: https://stackoverflow.com/questions/24211595/order-statistics-in-r
3. Fit distribution: https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.pdf