In All Likelihood - Wuji Shan

Wuji Shan

5/7/2022

Abstract

For this project, I decided to follow In All Likelihood track as my main focus. The first part of this report is the coding part for Exercise 4.25, 4.39, 4.27, and the Illinois Rain Analysis. Then I stated my learning during the working and the future plan between the coding and the reference part.

In All Likelihood

1. Exercise 4.25

Suppose U1,. . . , Un are an iid sample from the standard uniform distribution, and let U1,. . . , Un be the order statistics. Investigate the approximation median for n = 5 and n = 10.

```
f <- function(x, mu = 0, sigma = 1) dunif(x, mu, sigma)
F <- function(x, mu = 0, sigma = 1) punif(x, mu, sigma, lower.tail = FALSE)

# probability distributions of order statistics
integrand <- function(x, r, n){
  x * (1 - F(x))^(r - 1) * F(x)^(n - r) * f(x)
}

E <- function(r, n){
  (1/beta(r, n - r + 1)) * integrate(integrand, -Inf, Inf, r, n)$value
}

medianUi <- function(k, n){
  m <- (k - 1/3) / (n + 1/3)
  return(m)
}
```

For n = 5:

```
E(3, 5)
```

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

```
medianUi(3, 5)
```

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

For n = 10:

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

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

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

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

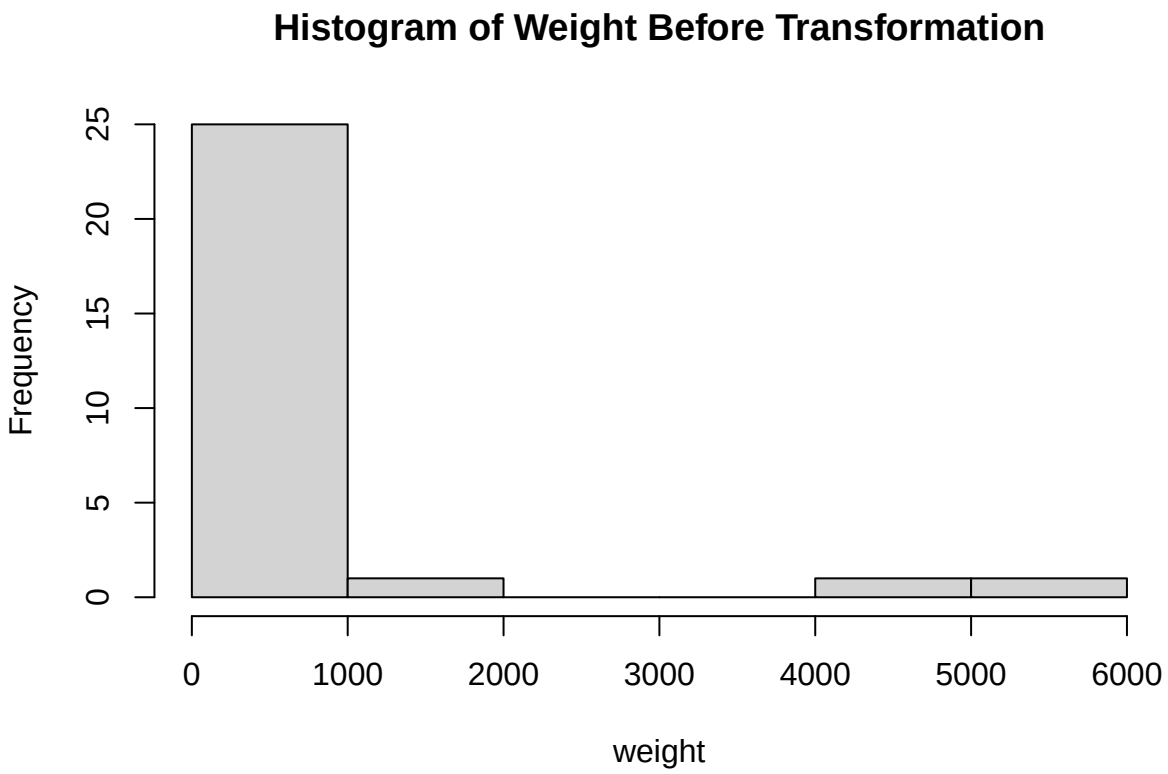
We can observe that the approximate median for n = 5 and n = 10 are same.

2. Exercise 4.39

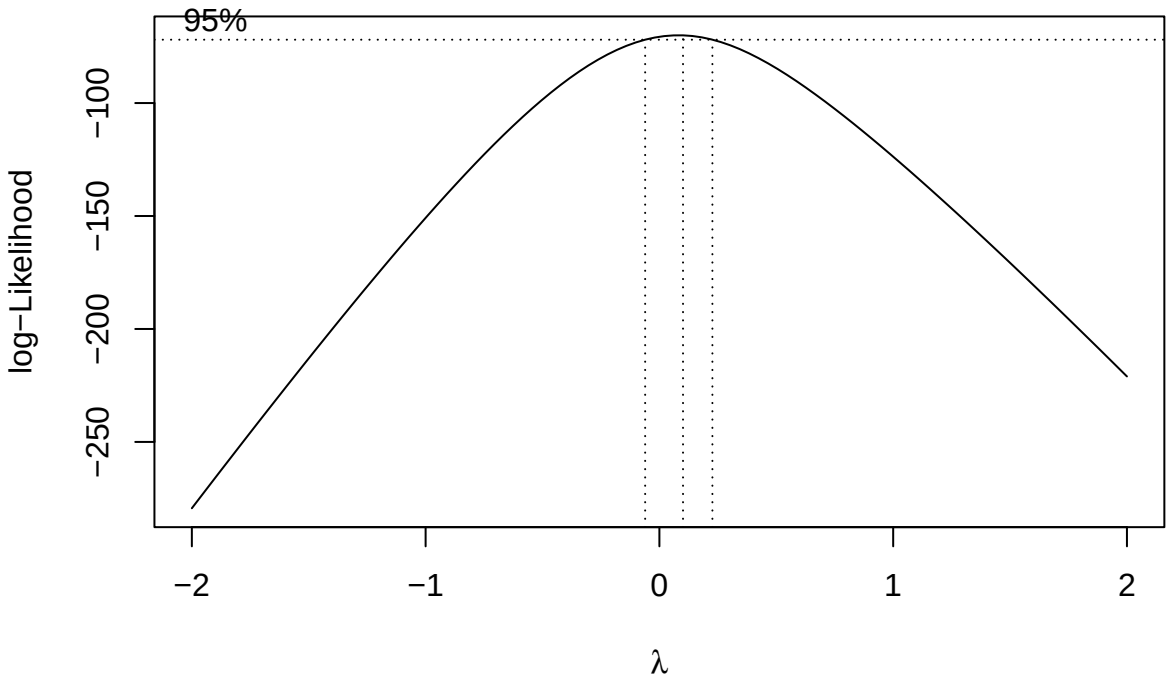
The data are the average adult weight (in kg) of 28 species of animal. Use the Box-Cox transformation family to find which transform would be sensible to analyse or present the data.

```
weight <- 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(weight, main = "Histogram of Weight Before Transformation")
```



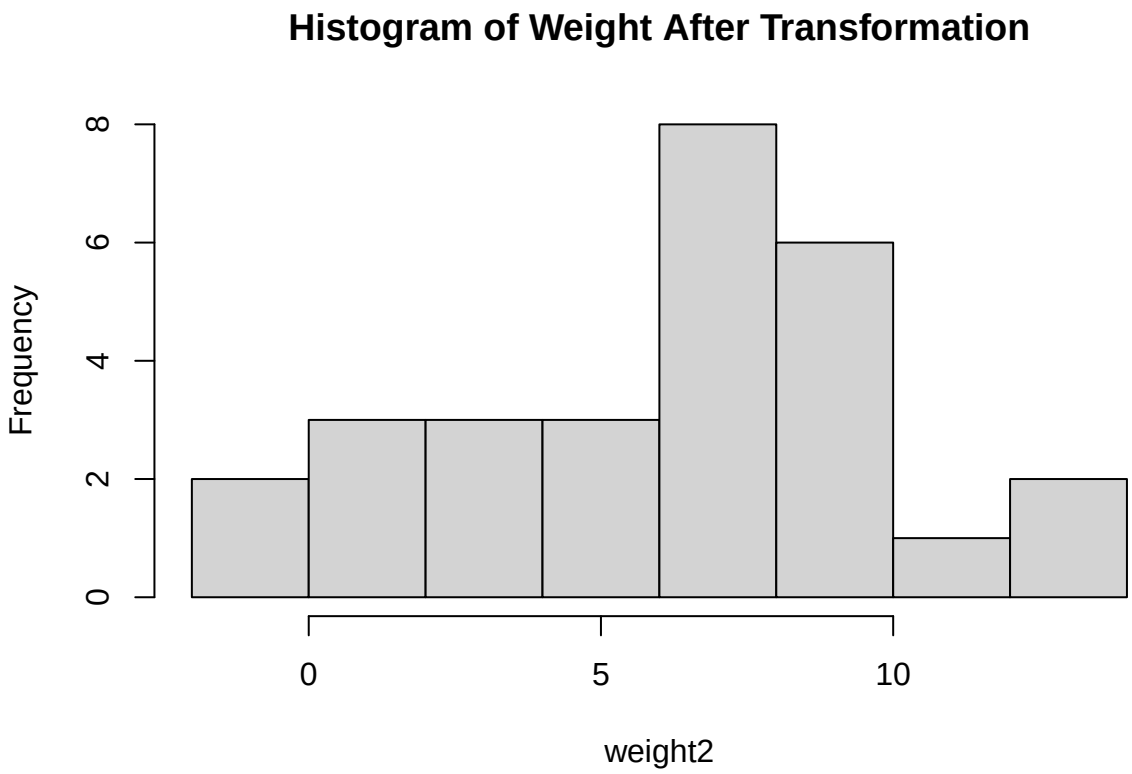

```
# boxcox transformation of the initial model
modeltran <- boxcox(lm(weight ~ 1))
```




```
# find the maximum of lambda
l <- modeltran$x[which.max(modeltran$y)]
l
```

```
## [1] 0.1010101
```

```
weight2 <- (weight ^ l - 1) / l
hist(weight2, main = "Histogram of Weight After Transformation")
```



3. Exercise 4.27

The data is the average amount of rainfall (in mm/hour) per storm in a series of storms in Valencia, southwest Ireland. Data from two months are reported below.

```
Jan <- 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)

July <- 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)

Compare the summary statistics for the two months.

```
summary(Jan)
```

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

```
##  0.1000  0.1875  0.4250  0.7196  0.9000  3.1700
```

```
summary(July)
```

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

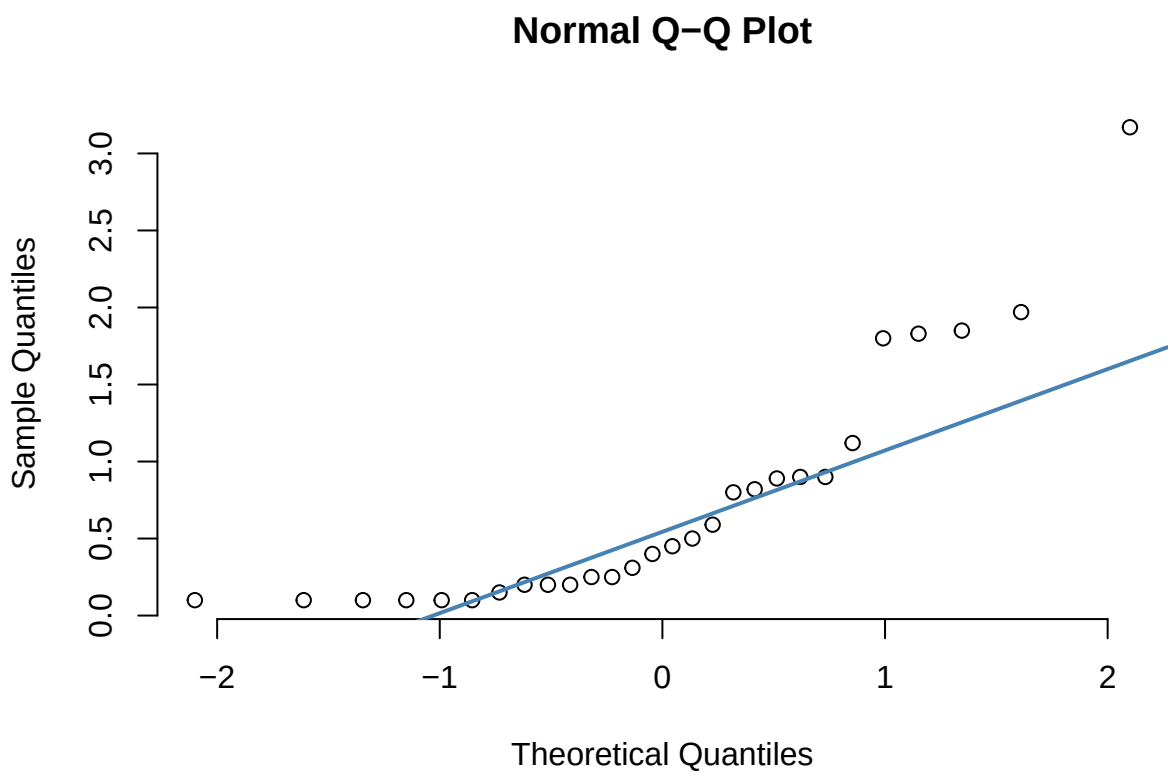
From the summary of two months, we can observe that January's statistics, except the minimum, are all higher than July's statistics.

(b)

Look at the QQ-plot of the data and, based on the shape, suggest what model is reasonable.

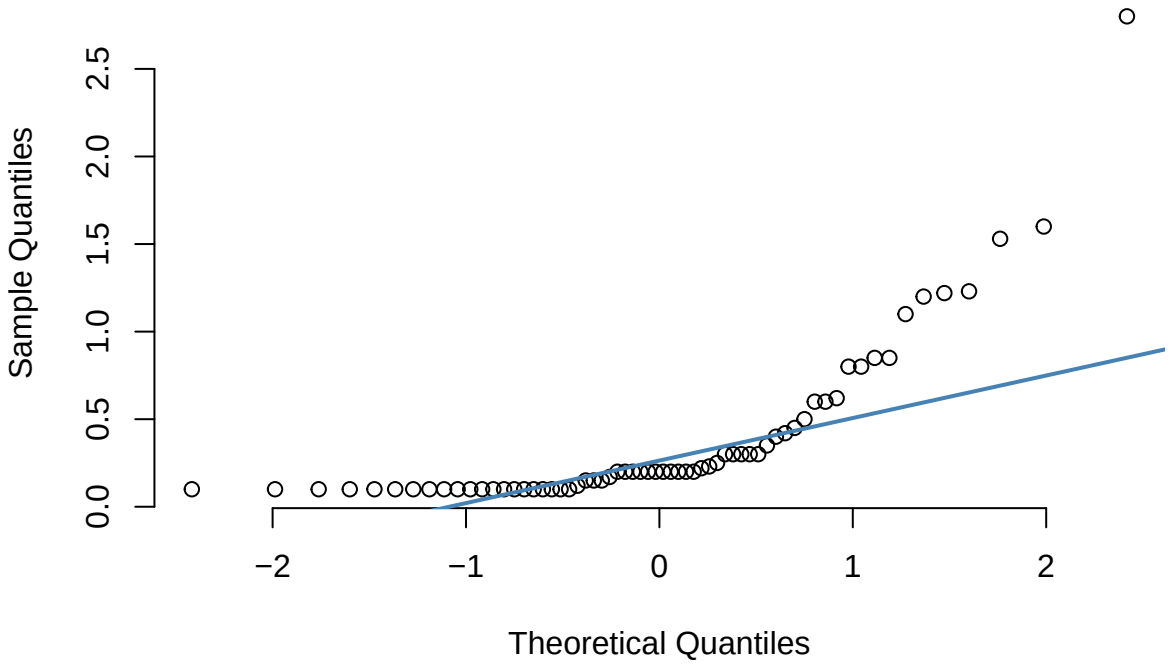
Here is the normal QQ-plot for January:

```
qqnorm(Jan, pch = 1, frame = FALSE)
qqline(Jan, col = "steelblue", lwd = 2)
```



Here is the normal QQ-plot for July:

```
qqnorm(July, pch = 1, frame = FALSE)
qqline(July, col = "steelblue", lwd = 2)
```



Based on the shape of QQ-plots, we can observe that the sample does not follow normal distribution. We may try gamma distribution model to fit the data later.

(c)

Fit a gamma model to the data from each month. Report the MLEs and standard errors, and draw the profile likelihoods for the mean parameters. Compare the parameters from the two months.

```
Jan_fit <- fitdist(Jan, "gamma", "mle")
Jan_fit
```

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

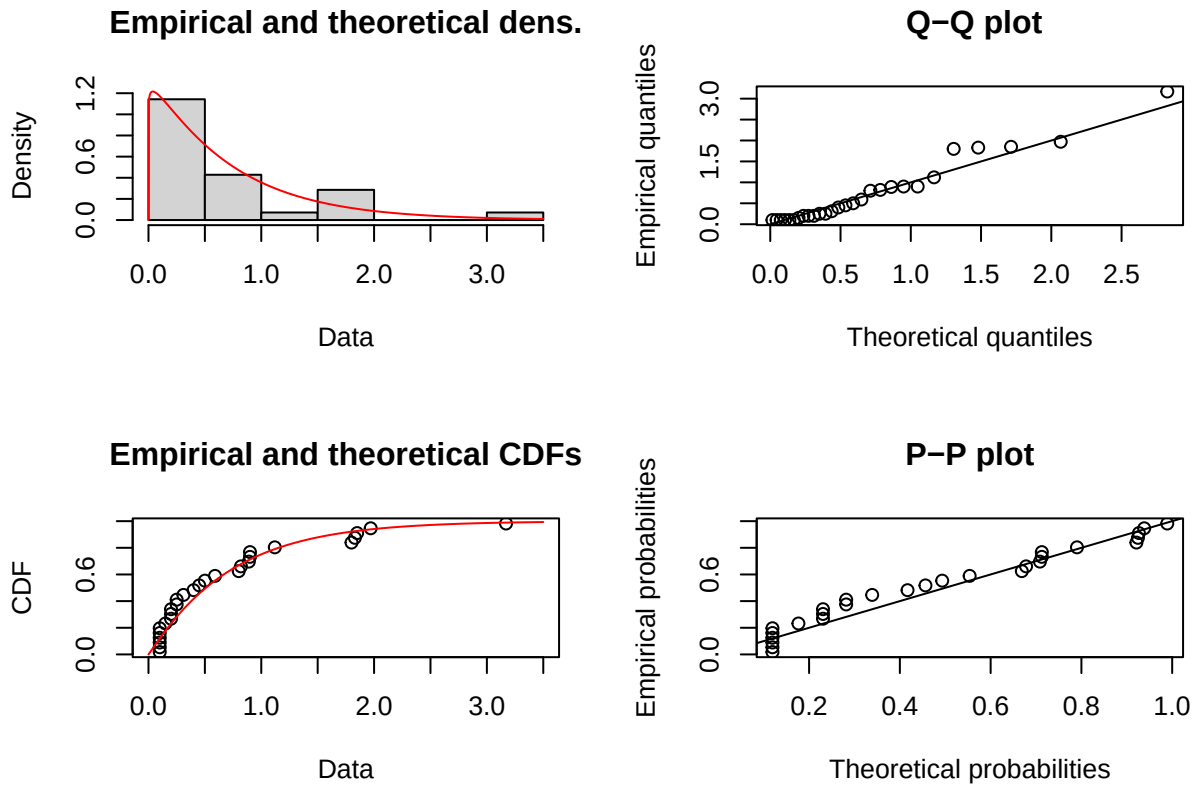
```
July_fit <- fitdist(July, "gamma", "mle")
July_fit
```

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

Since July's MLE is higher then that of January, July's model is better than January's.

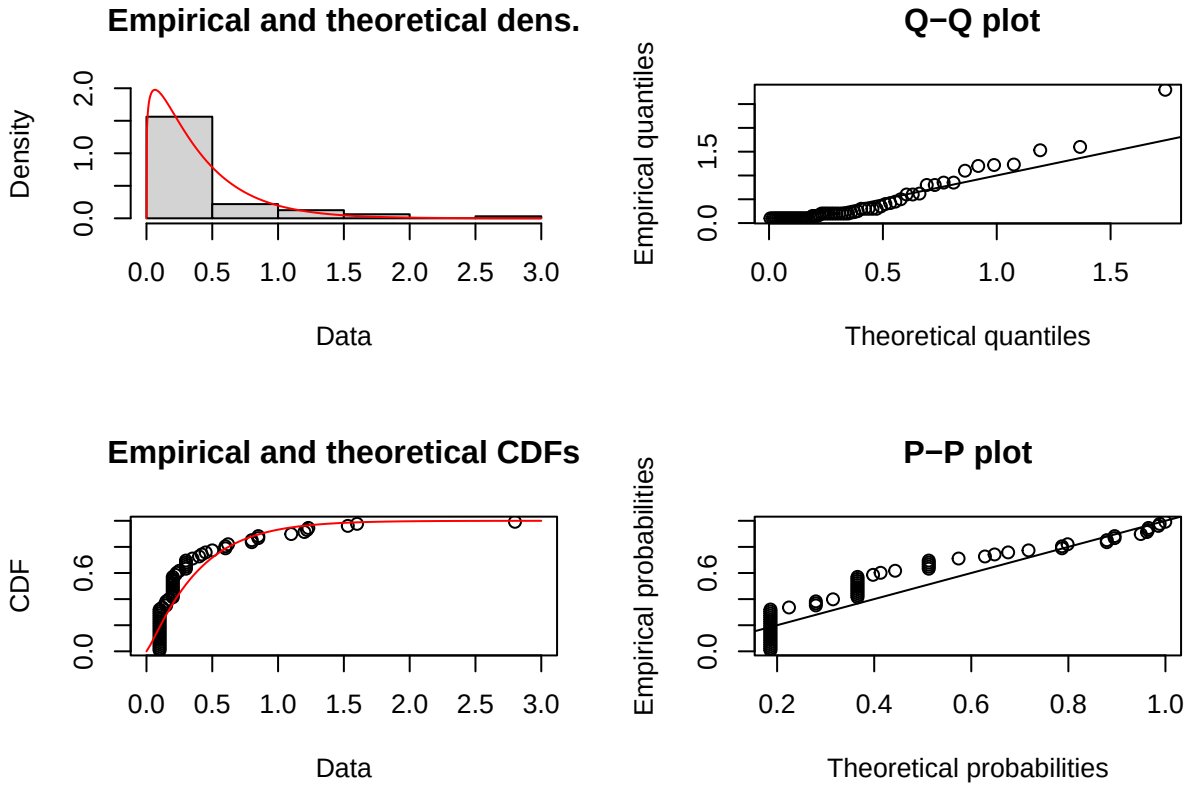
For Jan_fit:

```
plot(Jan_fit)
```



For July_fit:

```
plot(July_fit)
```




```
exp(Jan_fit$loglik)
```

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

```
exp(July_fit$loglik)
```

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

January's parameter is much lower than that of July.

(d)

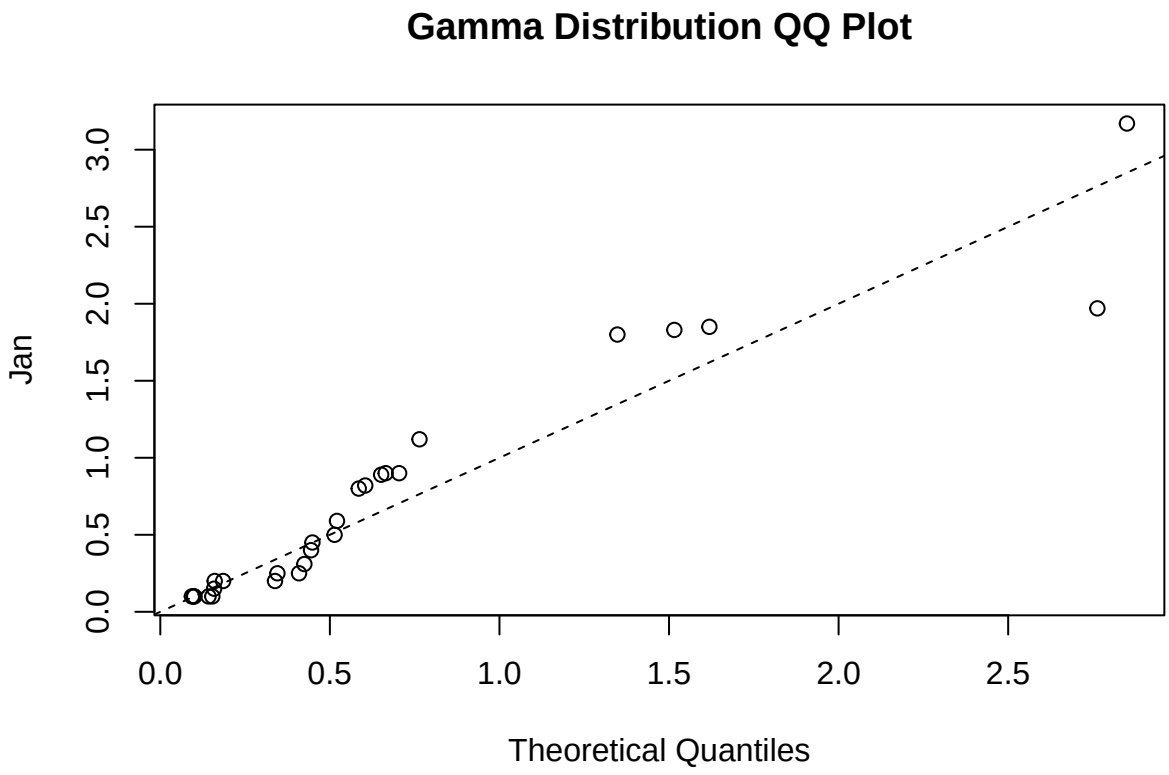
Check the adequacy of the gamma model using a gamma QQ-plot.

```
qqGamma <- function(x, ylab = deparse(substitute(x)),
                    xlab = "Theoretical Quantiles",
                    main = "Gamma Distribution QQ Plot",...)
{
    # Plot qq-plot for gamma distributed variable

    xx = x[!is.na(x)]
    aa = (mean(xx))^2 / var(xx)
    ss = var(xx) / mean(xx)
```

```
    test = rgamma(length(xx), shape = aa, scale = ss)

    qqplot(test, xx, xlab = xlab, ylab = ylab, main = main,...)
    abline(0,1, lty = 2)
}
```

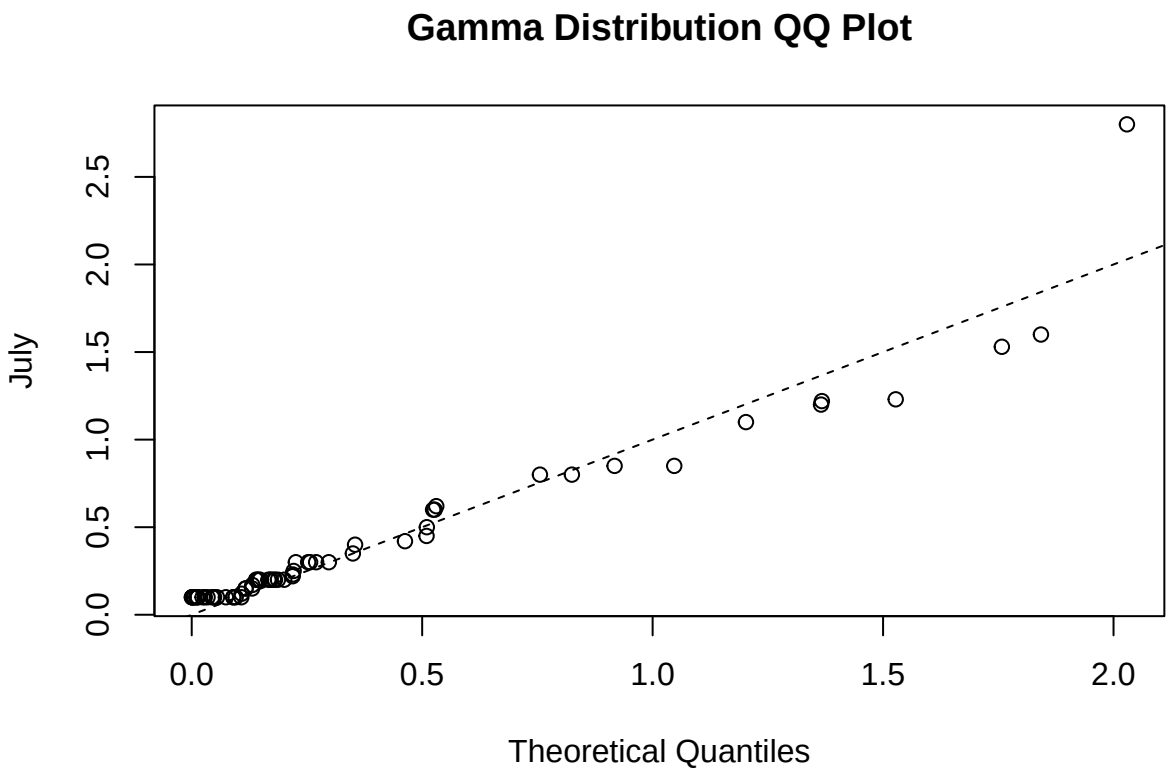
For January:

```
qqGamma(Jan)
```



For July:

```
qqGamma(July)
```



From the Gamma distribution QQ-plots, it seems that July's model is better.

4. Illinois Rain Analysis

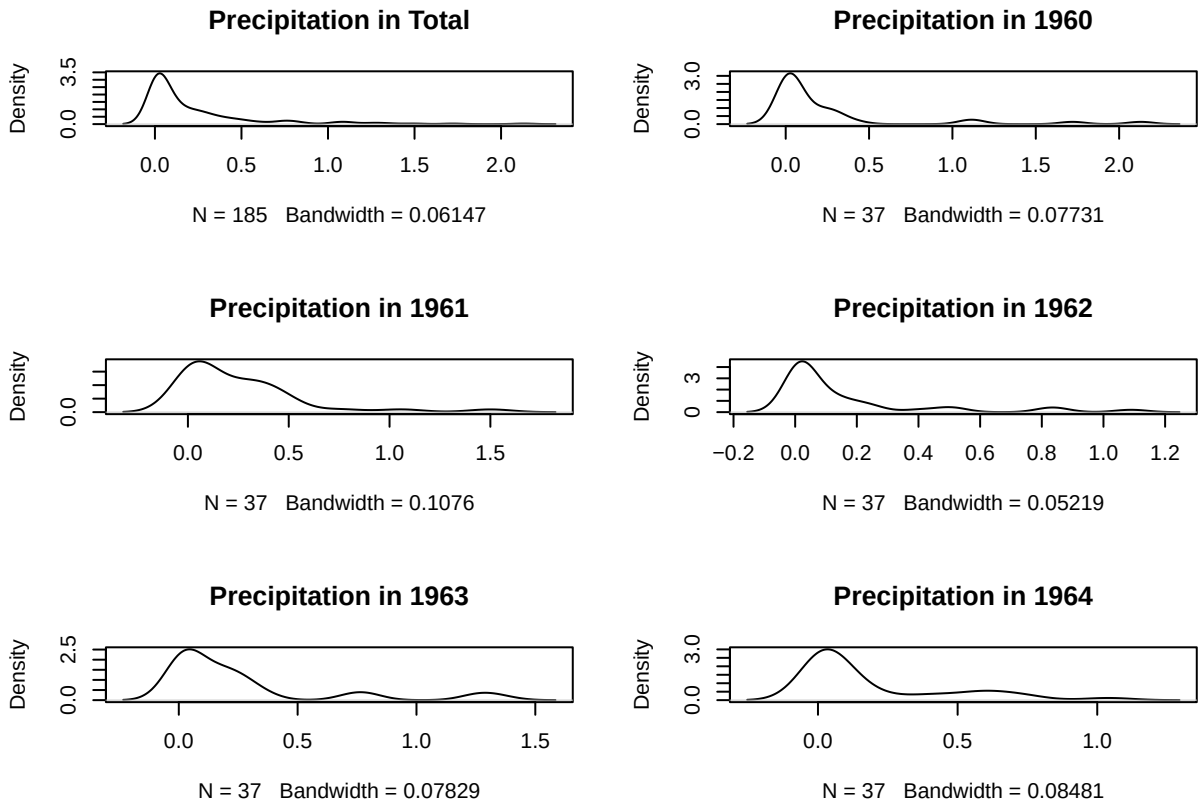
First Step

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.

```
# read the xlsx file
rain <- read.xlsx("Illinois_rain_1960-1964.xlsx")
```

The following are density plots for each year from 1960 to 1964 and in total.

```
# draw the density plot for each year and the total
par(mfrow = c(3, 2))

rain_nona <- na.omit(rain)

plot(density(unlist(rain_nona)), main = "Precipitation in Total")
plot(density(rain_nona$`1960`), main = "Precipitation in 1960")
plot(density(rain_nona$`1961`), main = "Precipitation in 1961")
plot(density(rain_nona$`1962`), main = "Precipitation in 1962")
plot(density(rain_nona$`1963`), main = "Precipitation in 1963")
plot(density(rain_nona$`1964`), main = "Precipitation in 1964")
```



Then I fit the data to check its distribution using MLE. The median and 95% confidence interval values are shown below.

```r
fit_rain <- fitdist(unlist(rain_nona), "gamma", method = "mle")
summary(bootdist(fit_rain))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median     2.5%     97.5%
## shape 0.4522745 0.386803 0.5375984
## rate  2.0313769 1.574937 2.7272101
```

Second Step

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?

```r
# calculate the mean value for the precipitation data in total
mean <- fit_rain$estimate[1] / fit_rain$estimate[2]
# calculate the mean value for each year
app <- apply(rain, 2, mean, na.rm = TRUE)

agg <- c(app, as.numeric(mean)) %>% round(5)
names(agg) <- c("1960", "1961", "1962", "1963", "1964", "mean")
agg
```

```
##    1960    1961    1962    1963    1964    mean
## 0.22029 0.27494 0.18475 0.26243 0.18711 0.22337
```

```r
storm <- c(nrow(rain) - apply(is.na(rain), 2, sum), '/')
data.frame(rbind(agg, storm))
```

```
##         X1960   X1961   X1962   X1963   X1964    mean
## agg   0.22029 0.27494 0.18475 0.26243 0.18711 0.22337
## storm      48      48      56      37      38       /
```

We can observe get some observations when comparing precipitation of each year to the mean value. 1961 and 1963 are wet years, 1962, 1960 and 1964 are dry years. Additionally, both of reasons, including more storms and individual storms produced more rain, have affected the amount of precipitation.

Third Step

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 believe the results of my analysis could be more generalizable if the data set includes data from a larger time range. I think the next steps after the analysis would be the data collection and applying the same analysis steps above to a large data set to check whether the results would change. Since I identified wet and dry years based on the comparison between mean value for each year and the total, if the data set includes data from more years, the mean value in total might change and then lead to the difference of wet and dry years identification results.

What I learned and Next Steps

For MA 677 this semester, I learned several significant statistical theories as the basis for the principles we apply in statistical practice. I learned a lot about how to describe and demonstrate the approaches used during working. Additionally, readings helped a lot because they includes examples and computation, which helped me to understand theories faster and apply them into practice. Next I will practice skills of not only computing but also presenting.

Reference:

1. Jin, Yuli. I learned from Yuli a lot because he shared several resources he thought useful to help understand the questions and strategies. Also, for each question, Yuli often tried different methods to solve the problem, which also gives me inspiration both for this project and the future study I believe.
2. https://stackoverflow.com/questions/24211595/order-statistics-in-r?msclkid=fd6683dac5671
3. https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r