MA 677 Final Project

Zening Ye

2022/05/03

In All Likelihood

Question 4.25

```
u <- function(x) dunif(x,0,1) # pdf of uniform distribution
U <- function(x) punif(x,0,1,lower.tail=F) # cdf of uniform distribution

# Probability distributions of order statistics
integrand <- function(x,r,n) {
  x * (1 - U(x))^(r-1) * U(x)^(n-r) * u(x)
}

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

approxmedi <- function(i,n) {
  m <- (i-1/3)/(n+1/3)
  return(m)
}

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

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

```
approxmedi(2.5,5)
```

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

```
# n = 10
(result(4,10) + result(5,10))/2
```

```
## [1] 0.4090909
```

```
(approxmedi(4,10) + approxmedi(5,10))/2
```

```
## [1] 0.4032258
```

Question 4.27

(a)

```
c1 <- data.frame(x=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))
c2 <- data.frame(x=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))
print(summary(c1))
```

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

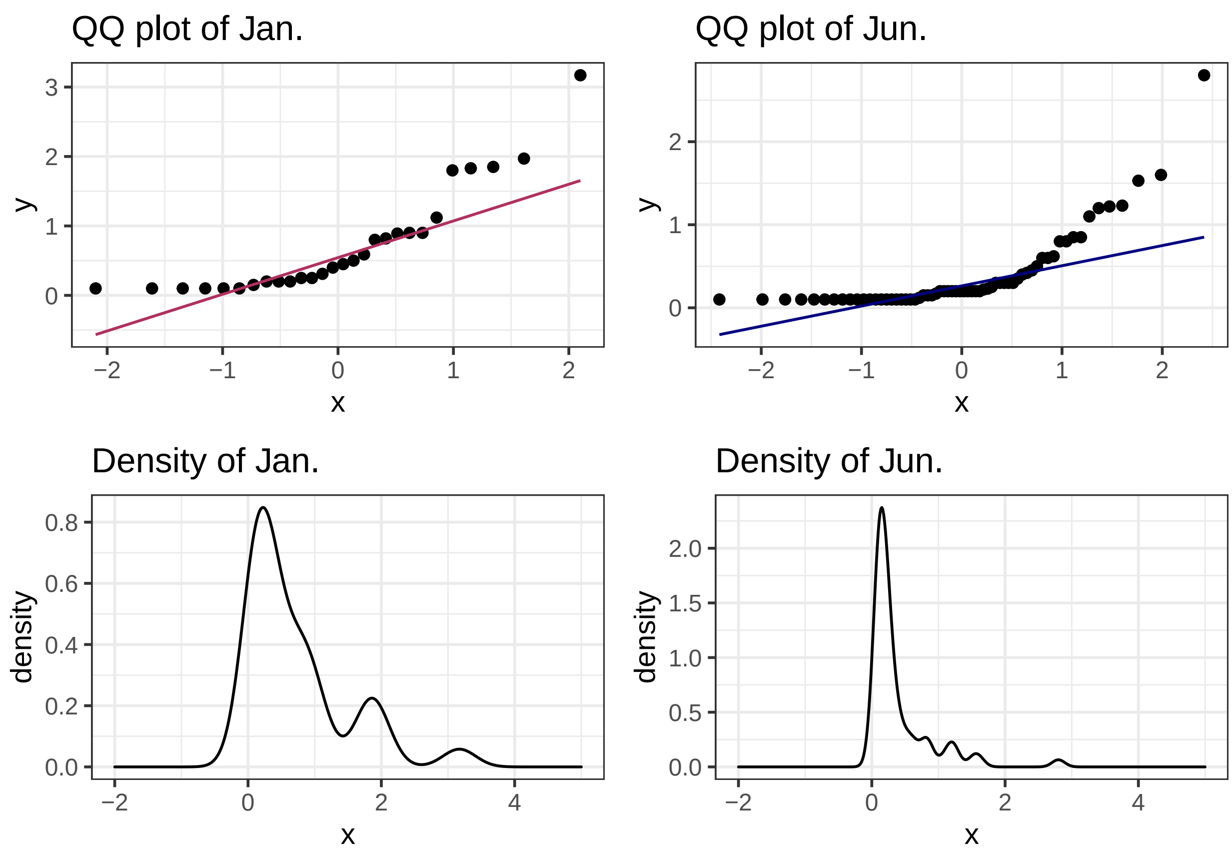
```
print(summary(c2))
```

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

According to the summary for both data, January mostly has higher value than June.

(b)

```
q1_27b <- ggplot(c1, aes(sample=x)) + stat_qq() + stat_qq_line(col='maroon') +
  labs(title='QQ plot of Jan.')
q2_27b <- ggplot(c2, aes(sample=x)) + stat_qq() + stat_qq_line(col='navy') +
  labs(title='QQ plot of Jun.')
h1_27b <- ggplot(c1) + geom_density(aes(x=x)) + xlim(c(-2,5)) +
  labs(title='Density of Jan.')
h2_27b <- ggplot(c2, aes(x=x)) + geom_density() +  xlim(c(-2,5)) +
  labs(title='Density of Jun.')
grid.arrange(q1_27b,q2_27b,h1_27b,h2_27b, ncol=2)
```



Based on qq plot, it looks both data are not follow normal distribution. However, the density plot indicates gamma distribution might fit with the model.

(c)

```r
# I used a simple way to conduct the gamma distribution -- fitdistrplus
# Reference: https://www.statology.org/fit-gamma-distribution-to-dataset-in-r/

jan <- fitdist(c1$x, distr = "gamma", method = "mle")
summary(jan) # summary of January
```

```
## 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 <- fitdist(c2$x, distr = "gamma", method = "mle")
summary(july) # summary of June
```

```
## 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
```

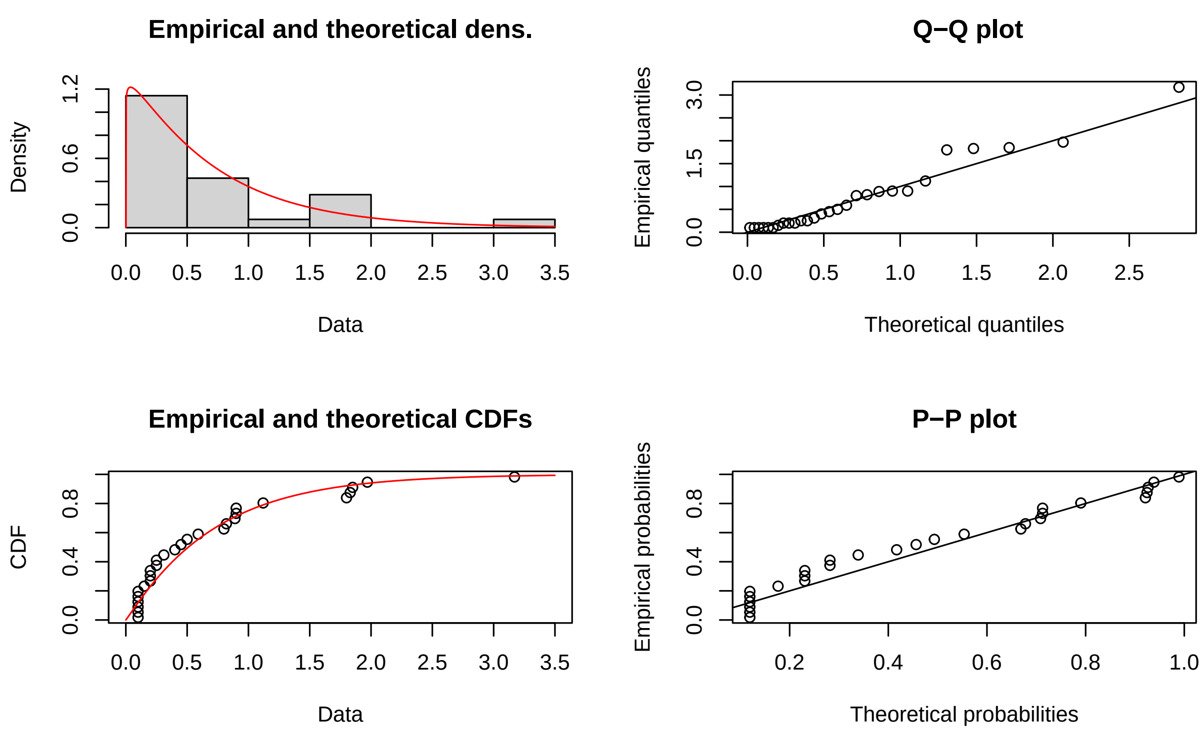
```
# MLE
exp(jan$loglik);exp(july$loglik)
```

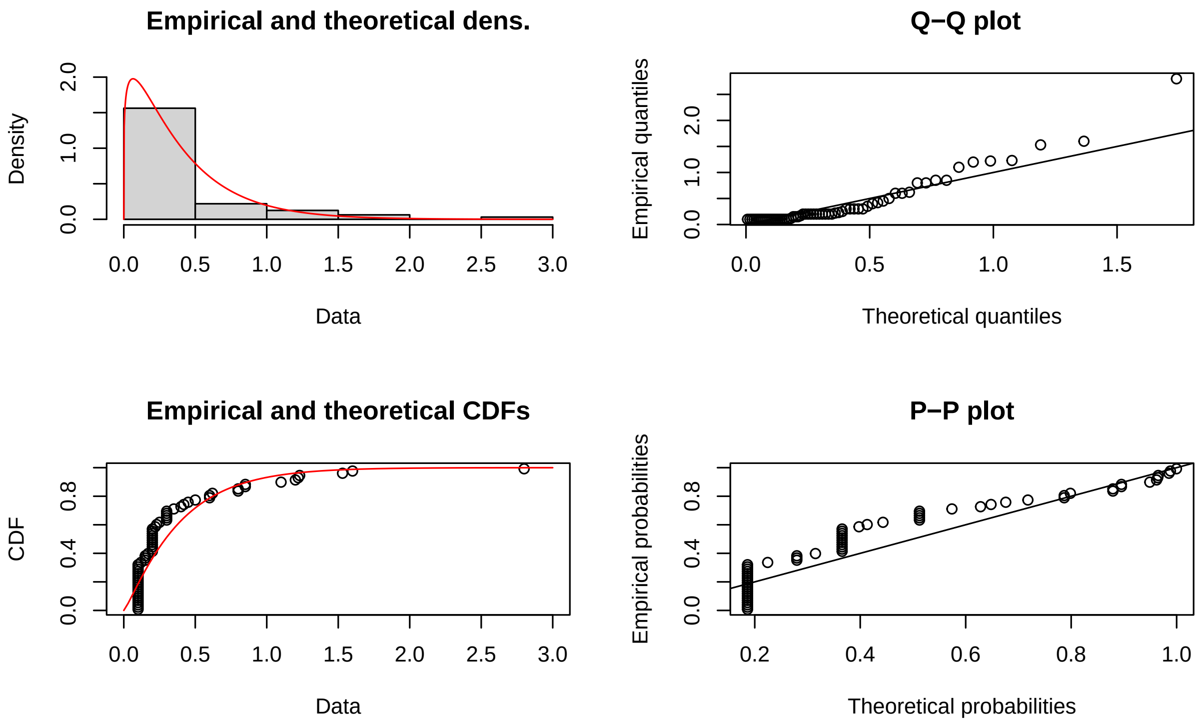
```
## [1] 7.11117e-09

## [1] 0.02638693
```

```
# sd
jan$sd;july$sd
```

```
##     shape      rate
## 0.2497495 0.4396202

##     shape      rate
## 0.1891196 0.5936302
```

```
# plot the result for both months
par(mfrow=c(1,2))
plot(jan);plot(july)
```

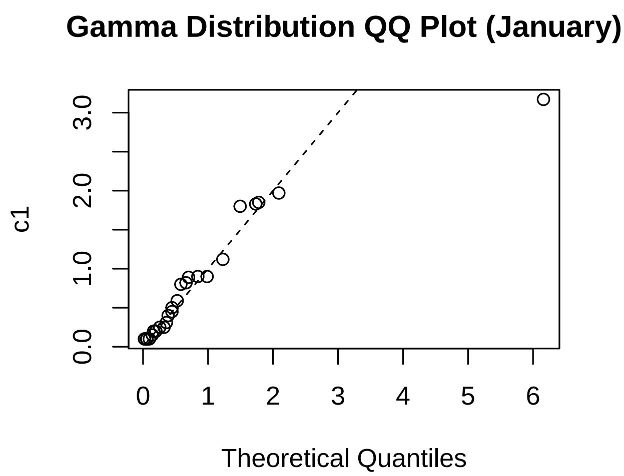


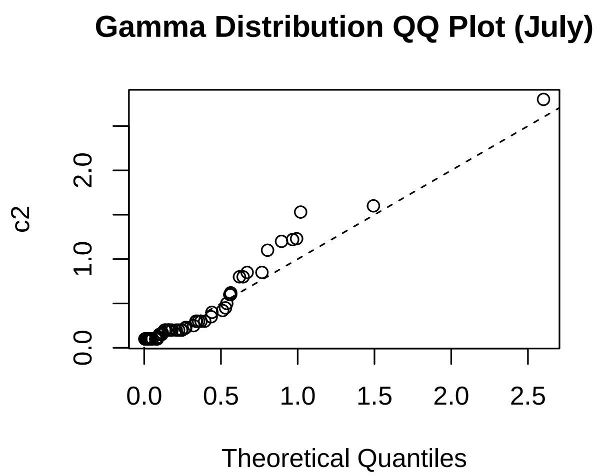


By using "fitdistrplus", I conducted two gamma distributions. The MLE for January data is 7.11117e-09, for June is 0.02638693. The standard error for both months I included before the plots of data. As you can see, the MLE of July is higher than January, which means the model of July is better.

(d)

```
# qq gamma plot
# Reference: https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r
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)
}

par(mfrow=c(1,2))
qqGamma(c1,main="Gamma Distribution QQ Plot (January)") # January
qqGamma(c2,main="Gamma Distribution QQ Plot (July)") # July
```

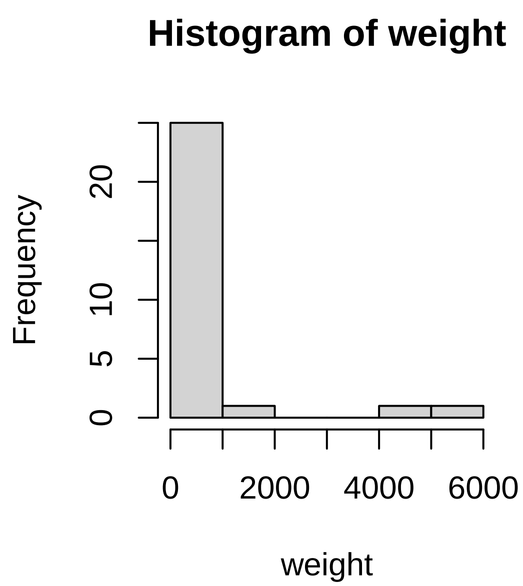


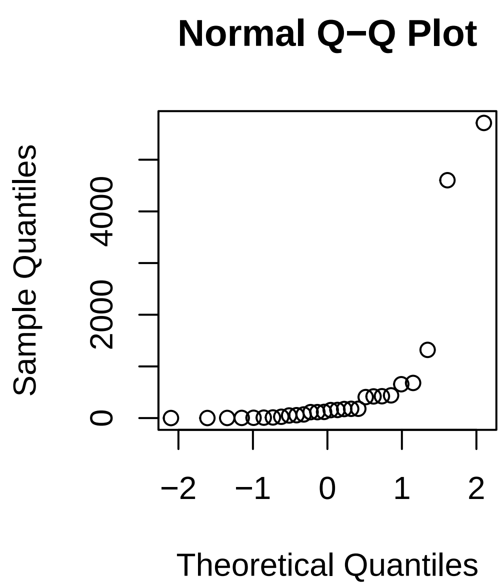


According to these plots, July might be better.

Question 4.39

```
# idea from https://r-coder.com/box-cox-transformation-r/
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)

par(mfrow=c(1,2))
hist(weight) # histogram of the data
qqnorm(weight) # checking the distribution
```

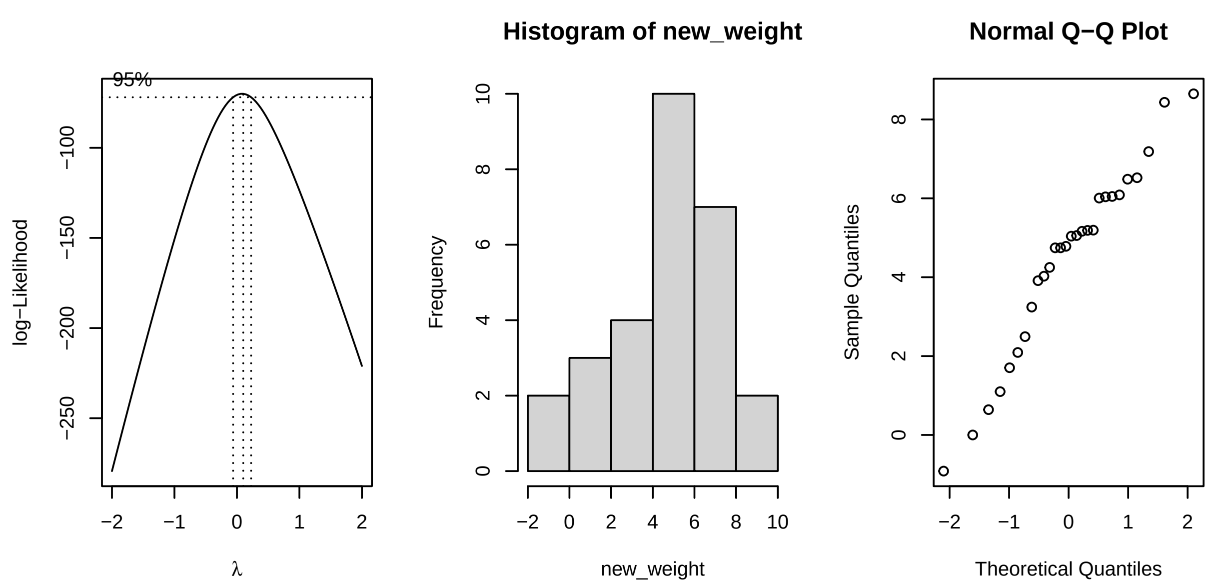




```r
# using linear regression to fit the data
og <- lm(weight ~ 1)
summary(og)
```

```
##
## Call:
## lm(formula = weight ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -574.1 -552.3 -437.5 -154.5 5137.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    574.5      252.3   2.277   0.0309 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1335 on 27 degrees of freedom
```

```r
# box-cox transformation
par(mfrow=c(1,3))
bc <- boxcox(og)

# Since the 0 is contain in 95% CI, we use log transformation
new_weight <- log(weight)

# plot the new weight distribution
hist(new_weight)
qqnorm(new_weight)
```



```r
# exact lambda
(lambda <- bc$x[which.max(bc$y)])
```

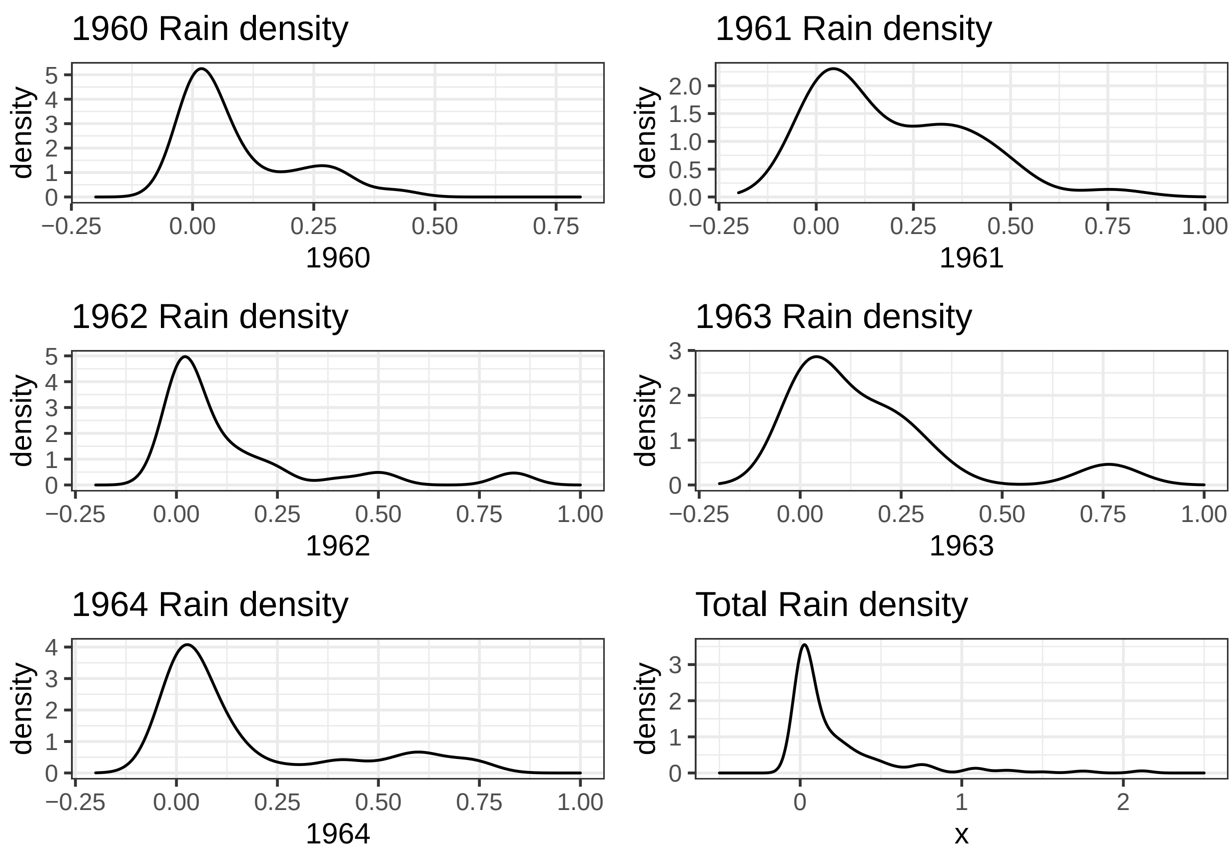
```
## [1] 0.1010101
```

Illiois Rainfall

```r
# import data
ill_rain <- read_xlsx('Illinois_rain_1960-1964.xlsx')
total <- data.frame(x=unlist(ill_rain)) %>% na.omit()
```

(a) 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 con dent you are about your identification of the distribution and the accuracy of your parameter estimates.

```r
# plot the density curve from 1960 to 1964
rain1960 <- ill_rain %>% na.omit(1960) %>% ggplot(aes(x=`1960`)) +
  geom_density() + labs(title='1960 Rain density') + xlim(c(-0.2,0.8))
rain1961 <- ill_rain %>% na.omit(1961) %>%ggplot(aes(x=`1961`)) +
  geom_density() + labs(title='1961 Rain density') + xlim(c(-0.2,1))
rain1962 <- ill_rain %>% na.omit(1962) %>% ggplot(aes(x=`1962`)) +
  geom_density() + labs(title='1962 Rain density') + xlim(c(-0.2,1))
rain1963 <- ill_rain %>% na.omit(1963) %>% ggplot(aes(x=`1963`)) +
  geom_density() + labs(title='1963 Rain density') + xlim(c(-0.2,1))
rain1964 <- ill_rain %>% na.omit(1964) %>% ggplot(aes(x=`1964`)) +
  geom_density() + labs(title='1964 Rain density') + xlim(c(-0.2,1))
rain_total <- total %>% ggplot(aes(x=x)) +
  geom_density() + labs(title='Total Rain density') + xlim(c(-0.5,2.5))
grid.arrange(rain1960,rain1961,rain1962,rain1963,
             rain1964, rain_total,ncol=2)
```



I used "fitdistrplus" to conduct different models to compare each other. Using MLE and MSE is a good start, both models will use gamma distribution:

```
# using entire dataset to fit the model
mle <- fitdist(total$x, distr='gamma',method='mle')
mse <- fitdist(total$x, distr='gamma',method='mse')

# To ensure our prediction, we need to find out the confidence interval for both models.
summary(bootdist(mle))
```

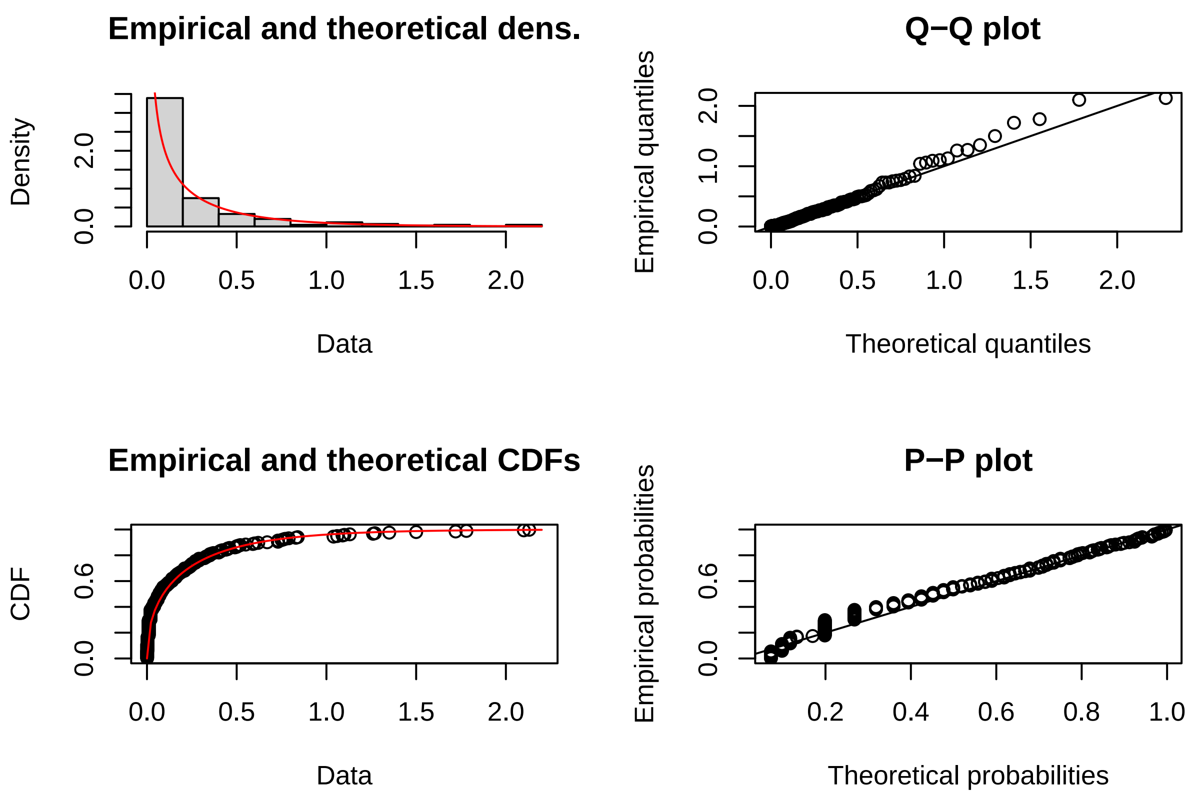
```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4439515 0.3840026 0.5177098
## rate  1.9878347 1.5868782 2.5370757
```

```
summary(bootdist(mse))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.7206567 0.6112711 0.8389103
## rate  1.3458511 1.0772319 1.6760579
```

According to the summary above, we noticed MLE's median and confidence interval is larger than the confidence interval in MSE, which means fitting model with gamma distribution and MLE is better for rain data. In addition, I made some plots for MLE method.

```
plot(mle)
```



As you can see, almost all the points are lying on the line in qq plot, empirical cdf plot and pp plot.

(b) 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?

```
mean_each <- apply(ill_rain,2,mean,na.rm=T) %>% round(digits = 4) # mean of each year
mean_all <- mle$estimate[1]/mle$estimate[2] # mean of all year
names(mean_all) <- 'total'
# combine data
tmean <- c(mean_each,mean_all) %>% round(4)

# count the storm number
num_storm <- c(apply(!is.na(ill_rain),2,sum),sum(apply(!is.na(ill_rain),2,sum))) %>%
  as.character()

result <- rbind(tmean,num_storm)
rownames(result) <- c('mean','num_storm')
kable(result) %>% kable_styling(full_width = T,position = 'center')
```

	1960	1961	1962	1963	1964	total
mean	0.2203	0.2749	0.1848	0.2624	0.1871	0.2244
num_storm	48	48	56	37	38	227

By using the table above to compare the mean between different year, we noticed 1962 and 1964 might be the dry year and for the rest of year, 1960,1961,1963, might be the wet year. Furthermore, we also concluded that the number of storm might not have a lot of relation with rainfall rate, for instance 1962. Therefore, there is a limit of influence between rainfall and storm.

(c) 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?

In my perspective, even though we conducted a good prediction from the model, it is not efficiency since we only have five years data. For the next step, I believe collecting more data to conduct a more complexity model is a plausible movement. Floyd Huff only conducted the theoretical part, he did not build a reliable model based on his thesis.

What I learned?

This semester I think I learned some statistical concepts that I had not leaned before, such as empirical statistic, statistic inference, order statistics, etc. This knowledge is challenging for a non-statistics student, but it was a very valuable experience. Next I think I will focus on learning some statistical methods such as stochastic gradient descent, and consolidate my programming skills to prepare myself.

Reference

- **Yuli Jin** for particular help
- "fitdistrplus" Package: https://www.statology.org/fit-gamma-distribution-to-dataset-in-r/
- "Box-Cox Transformation": https://r-coder.com/box-cox-transformation-r/
- "QQ Gamma Plot": https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r