final_project

Xihao Cao

2022/5/03

Exerciese 4.25

```
# we firstly define the distribution of the standard uniform distribution
f <- function(x, a=0, b=1) dunif(x, a,b)
F <- function(x, a=0, b=1) punif(x, a,b, lower.tail=FALSE)


# then we define the order statistics distribution in exercise 2.4 with
# expectation estimation function and median estimation function
orderStat <- function(x,r,n) {
  return(x * (1 - F(x))^(r-1) * F(x)^(n-r) * f(x))
}


expect <- function(r,n) {
  return((1/beta(r,n-r+1)) * integrate(orderStat,-Inf,Inf, r, n)$value)
}


median<-function(i,n){
  return((i-1/3)/(n+1/3))
}
# then calculate and compare n = 5 and n = 10 in the next block
```

the difference for n = 5 is the following which is very small, which means the approximation is great when n = 5.

```
expect(2.5,5) - median(2.5,5)
```

```
## [1] 0.01041666
```

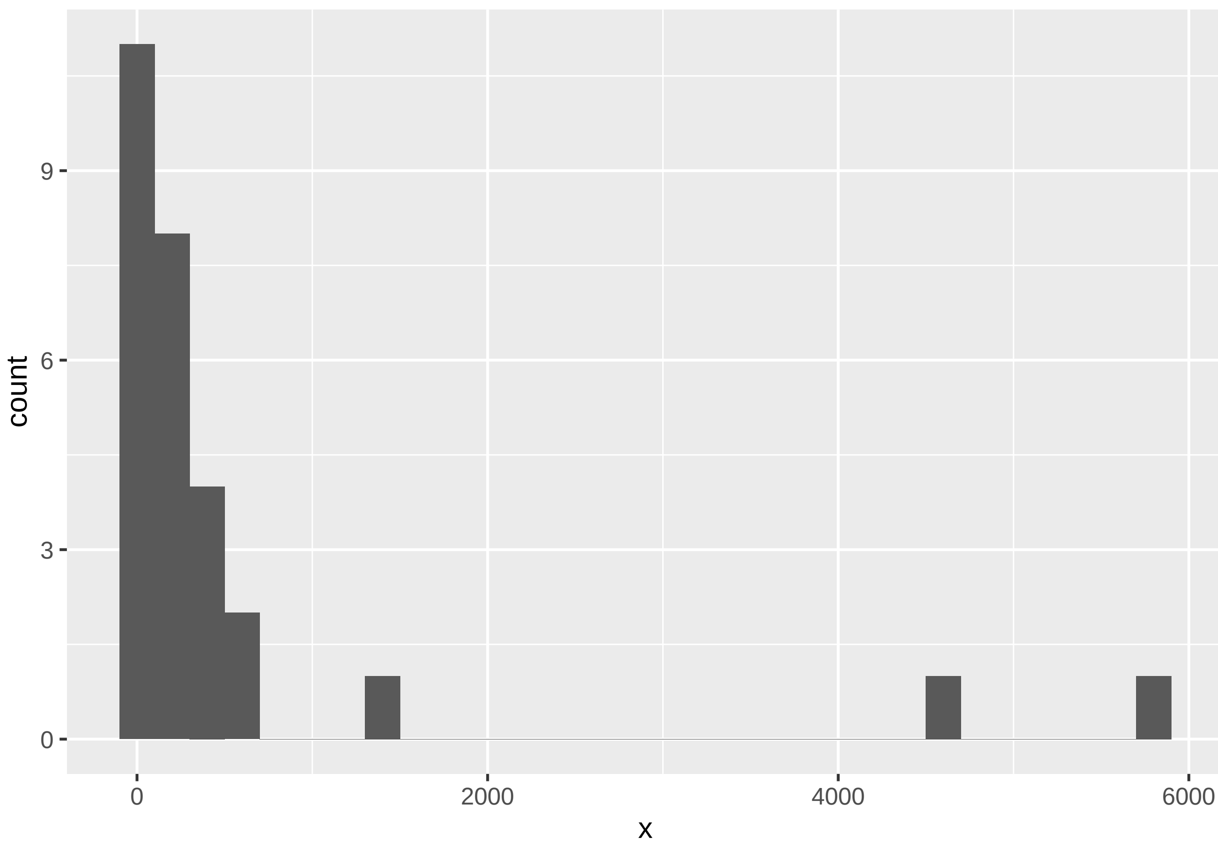
the difference for n = 10 is the following which is very small, which means the approximation is great when n = 10.

```
expect(5,10) - median(5,10)
```

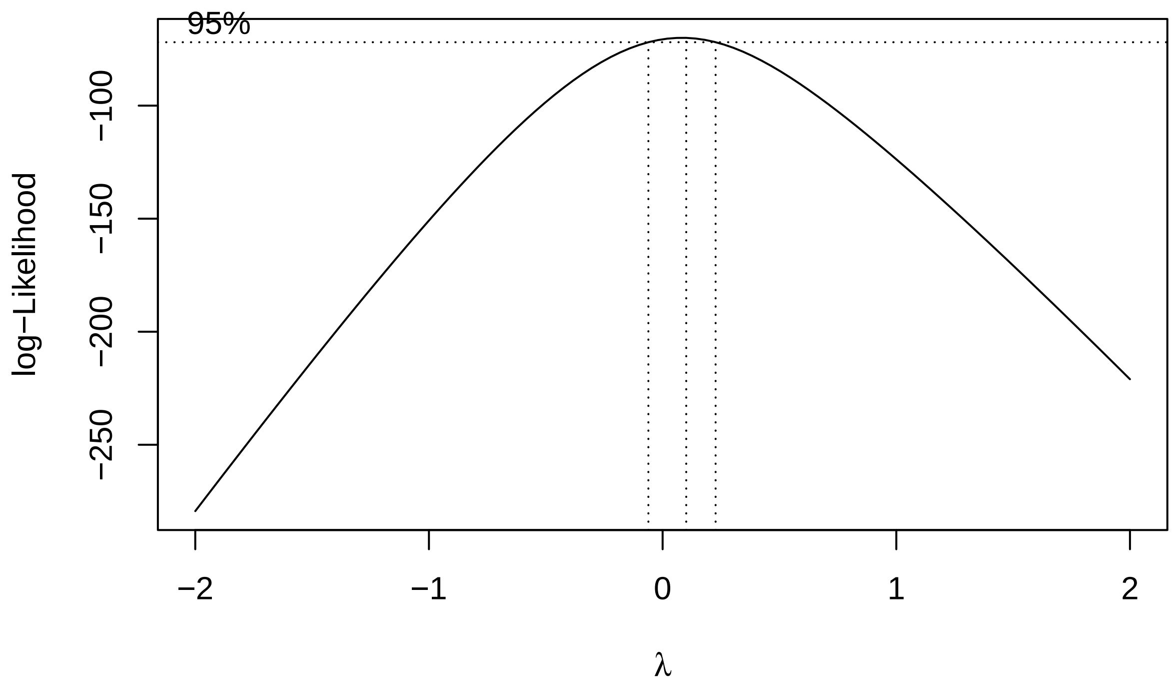
```
## [1] 0.002932548
```

Exercise 4.39

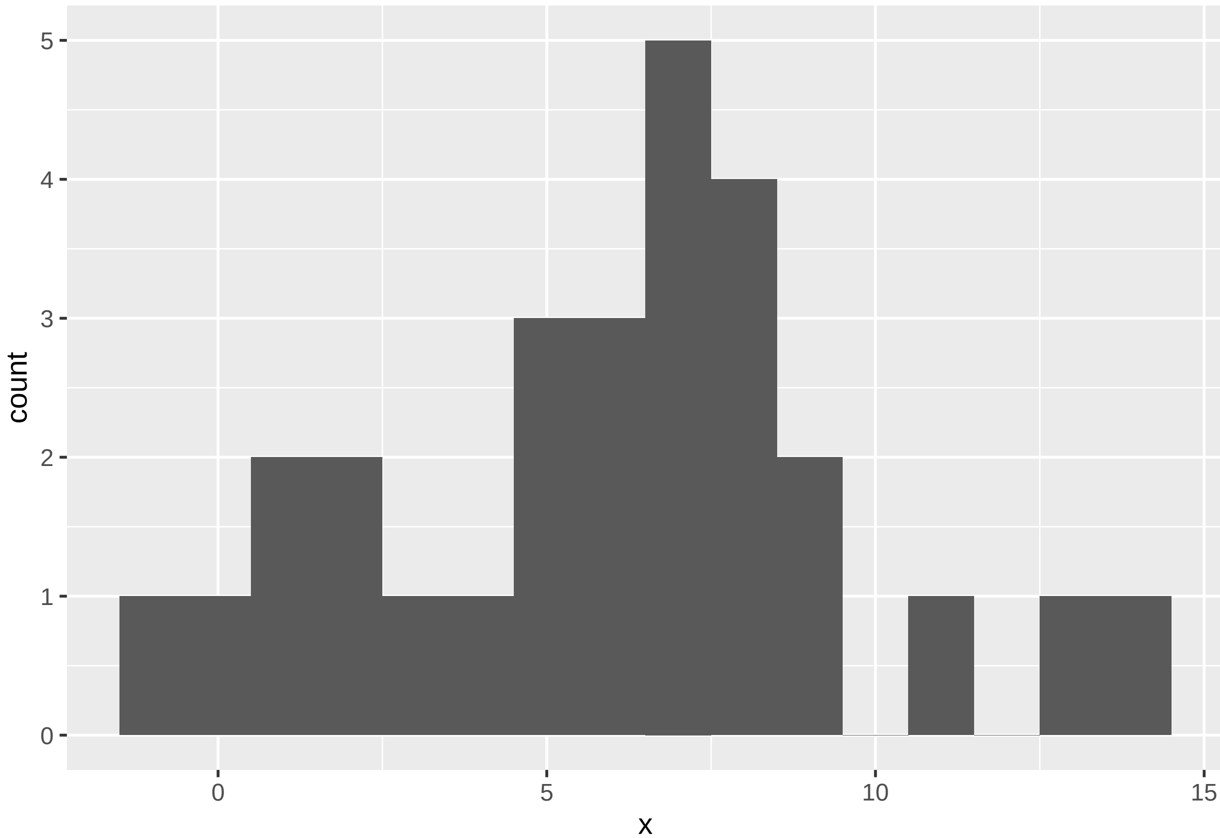
```
data<-data.frame(x = c(0.4,1.0,1.9,3.0,5.5,8.1,12.1,25.6,50.0,56.0,70.0,115.0,
                       115.0,119.5,154.5,157.0,175.0,179.0,180.0,406.0, 419.0,
                       423.0, 440.0, 655.0, 680.0, 1320.0, 4603.0, 5712.0))
ggplot(data) + geom_histogram(aes(x = x), binwidth = 200)
```




```
# make Boxcox transformation of the data
box <- boxcox(lm(data$x ~ 1))
```




```
# extract lambda
lambda <- box$x[which.max(box$y)]
tran_data <- (data ^ lambda - 1) / lambda
ggplot(tran_data) + geom_histogram(aes(x = x), binwidth = 1)
```



the first plot shows that the initial distribution of the data is not normally distributed. Then we conduct the Box–Cox transformation, and plot the lambda graph and extract the optimal lambda, which is 0.1010101. Then use the optimal lambda to transform the original data, now the new data is nearly normally distributed.

Exercise 4.27

Part a

```
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)


Jul<-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)

summary(Jan)
```

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

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

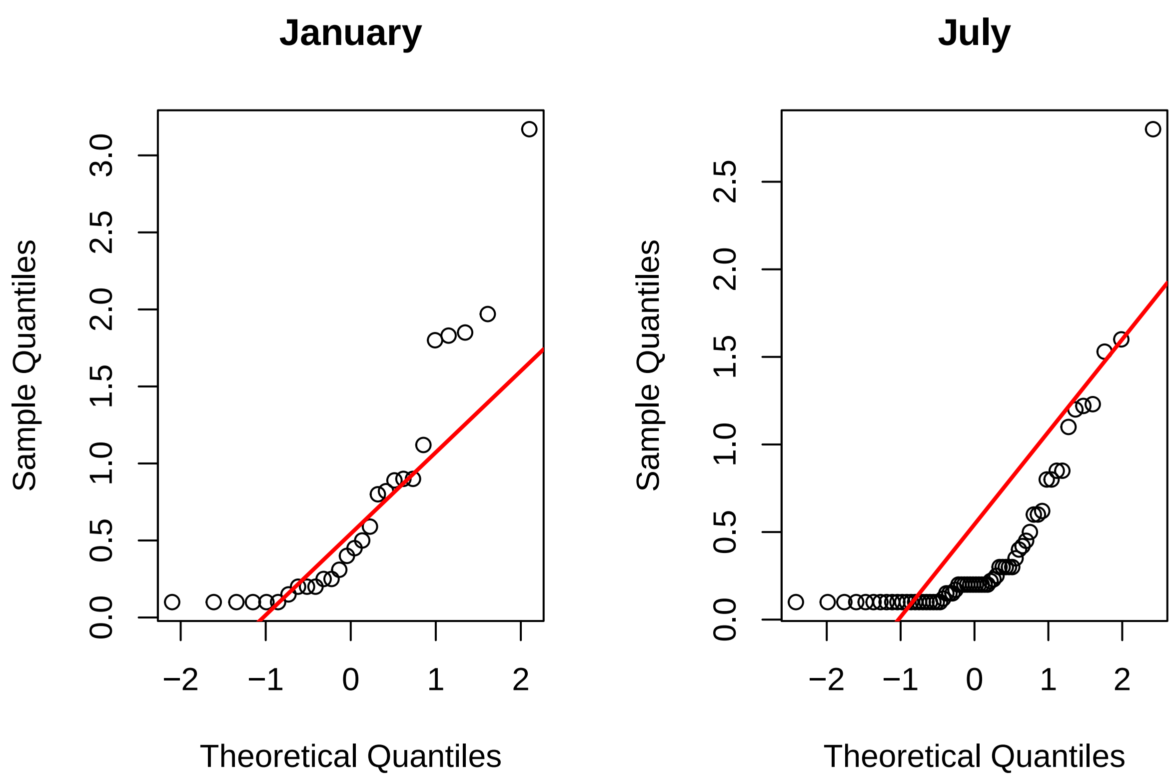
```r
summary(Jul)
```

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

The two collections of data share the same minimum value, but the 1st quantile, median, mean, 3rd quantile, and max value of the January data are all bigger than the data of July.

Part b

```r
par(mfrow = c(1, 2))
qqnorm(Jan, pch = 1, main = "January")
qqline(Jan, col = "red", lwd = 2)
qqnorm(Jul, pch = 1, main = "July")
qqline(Jan, col = "red", lwd = 2)
```



From the QQ plots, we can see that two collections of data do not follow the normal distribution. By the hint of this problem, I believe gamma distribution may work.

part c

```r
Jan.fit <-  fitdist(Jan,'gamma','mle')
July.fit <-  fitdist(Jul,'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
```

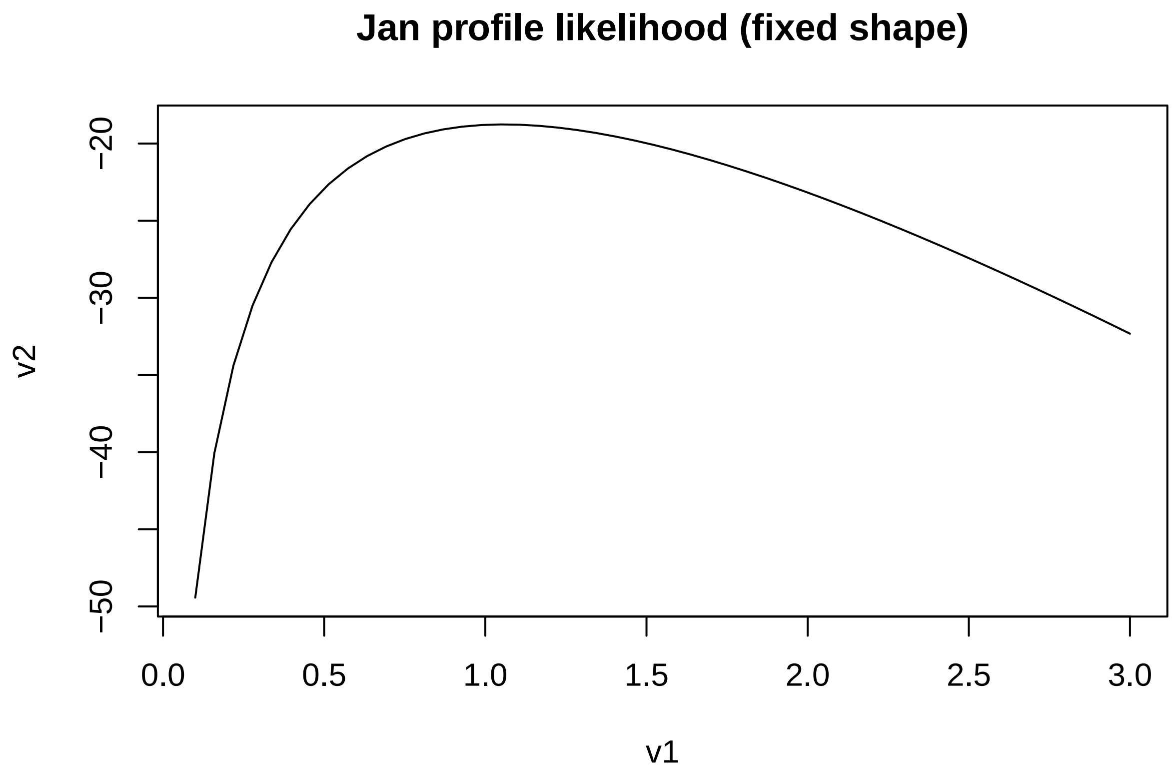
```r
summary(July.fit)
```

```
## 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
```

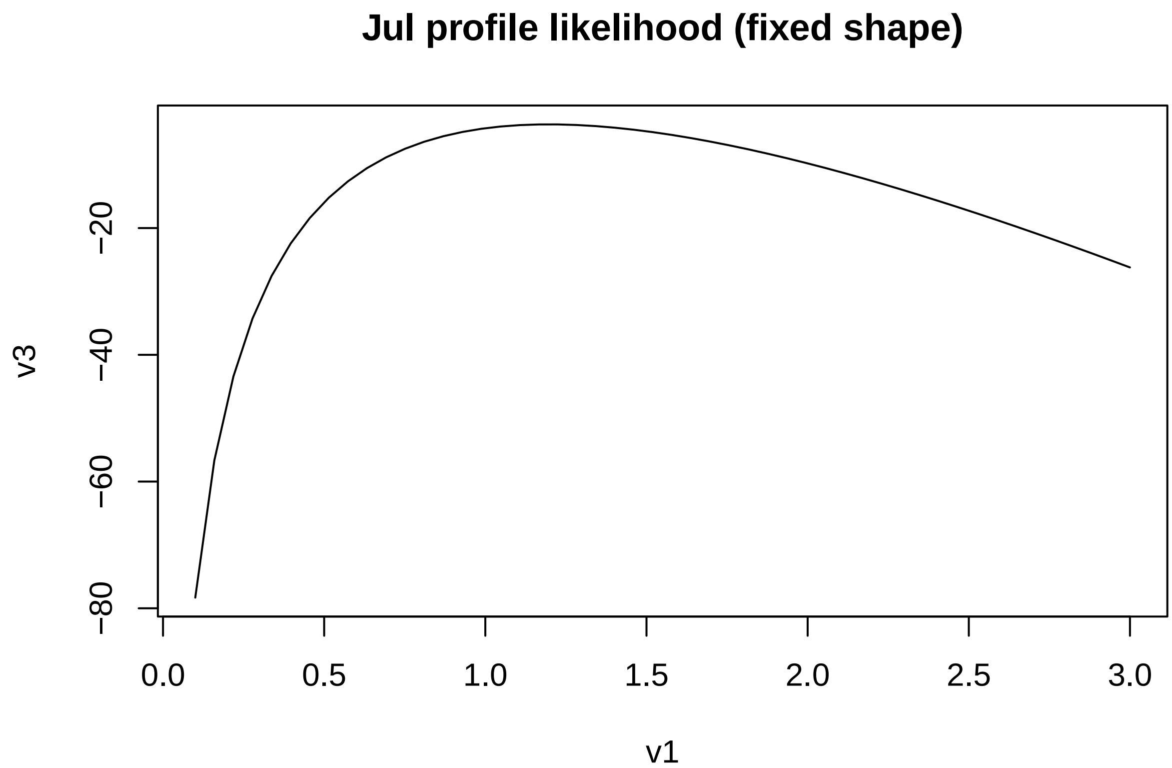
By reading the summary of two gamma models. We can see that the MLE of July's model is higher than the one of January's. We can also compare other statistics by their tables. Then we plot the profile likelihood of both months with the fixed shape.

```r
x <- Jan
prof_log_lik=function(a){
   b=(optim(1,function(z) -sum(log(dgamma(x,a,z)))))$par
   return(-sum(log(dgamma(x,a,b))))
 }

v1 <- seq(.1, 3, length=50)
v2 <- -Vectorize(prof_log_lik)(v1)
plot(v1, v2, type="l", main = 'Jan profile likelihood (fixed shape)')
```

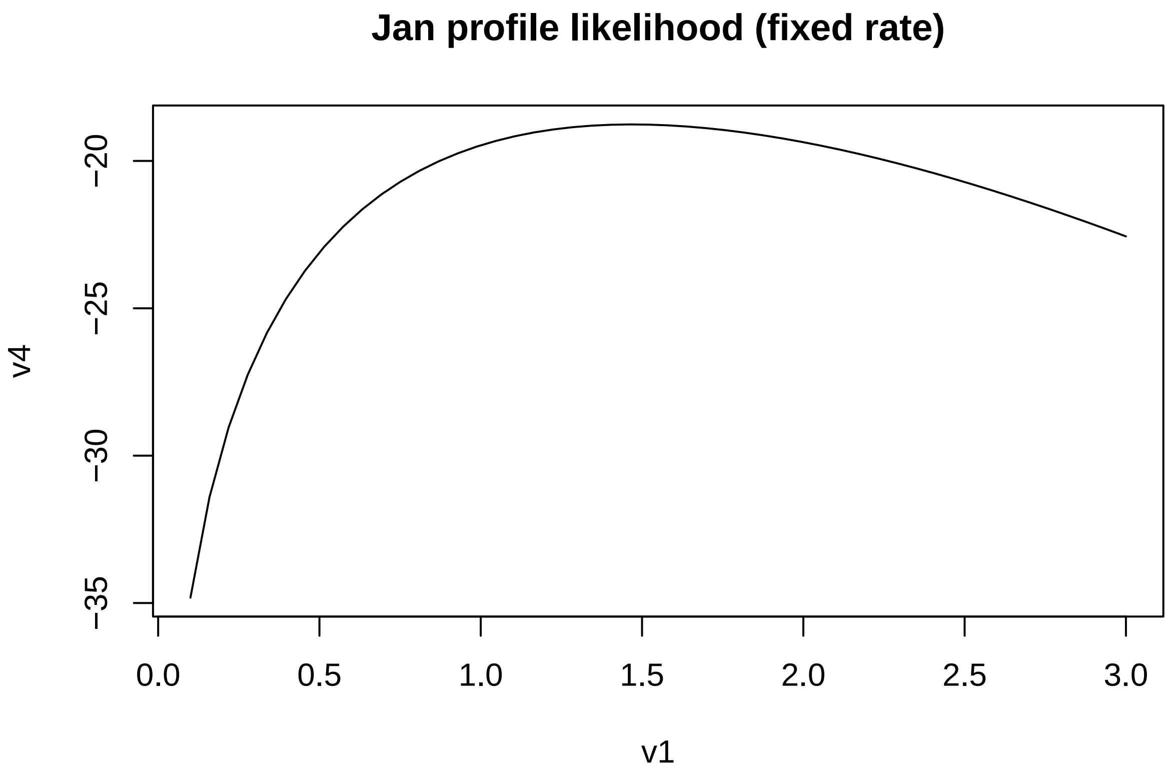



```r
x <- Jul
v3 <-  -Vectorize(prof_log_lik)(v1)
plot(v1, v3, type="l", main='Jul profile likelihood (fixed shape)')
```

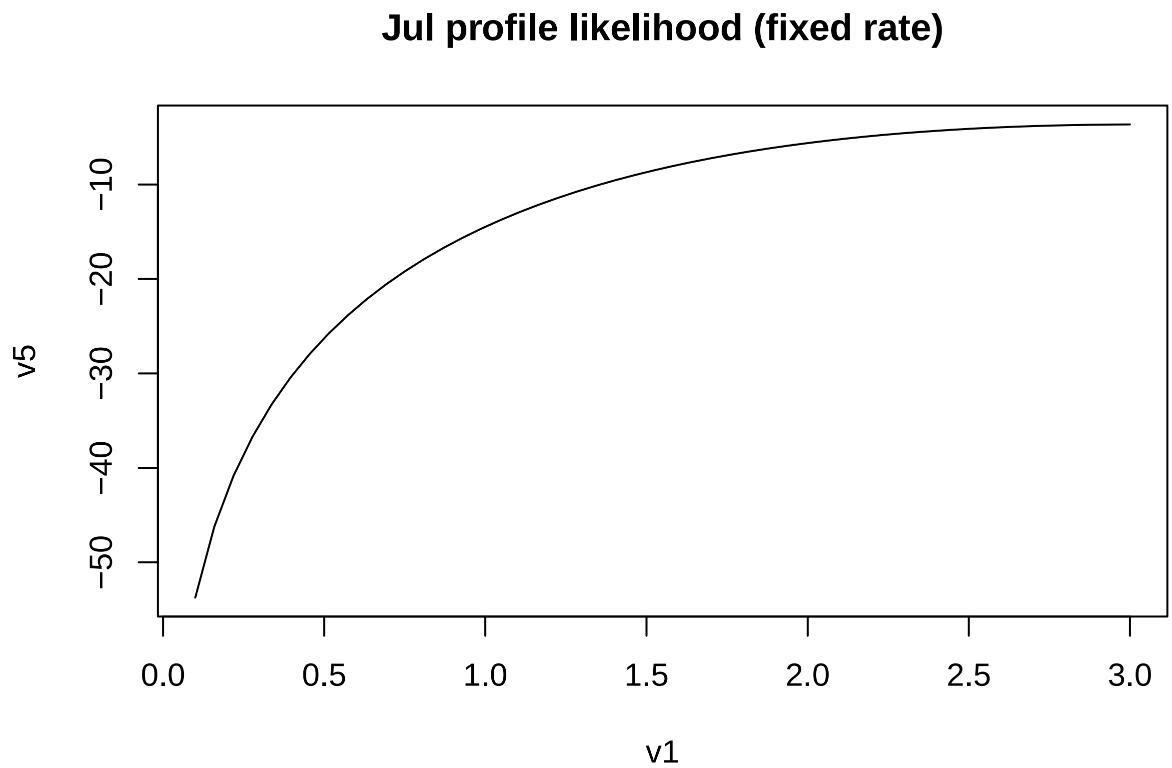


Then plot the profile likelihood of both mouth with the fixed rate.

```
x <- Jan
prof_log_lik=function(z){
   a=(optim(1,function(a) -sum(log(dgamma(x,a,z)))))$par
   return(-sum(log(dgamma(x,a,z))))
 }
v4 <- -Vectorize(prof_log_lik)(v1)
plot(v1, v4, type="l",main='Jan profile likelihood (fixed rate)')
```



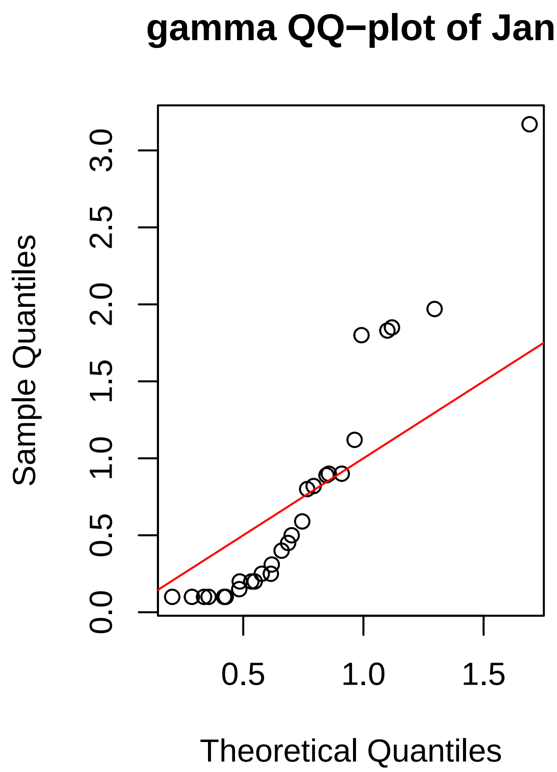

```r
x <- Jul
v5 <- -Vectorize(prof_log_lik)(v1)
plot(v1, v5, type="l",main='Jul profile likelihood (fixed rate)')
```

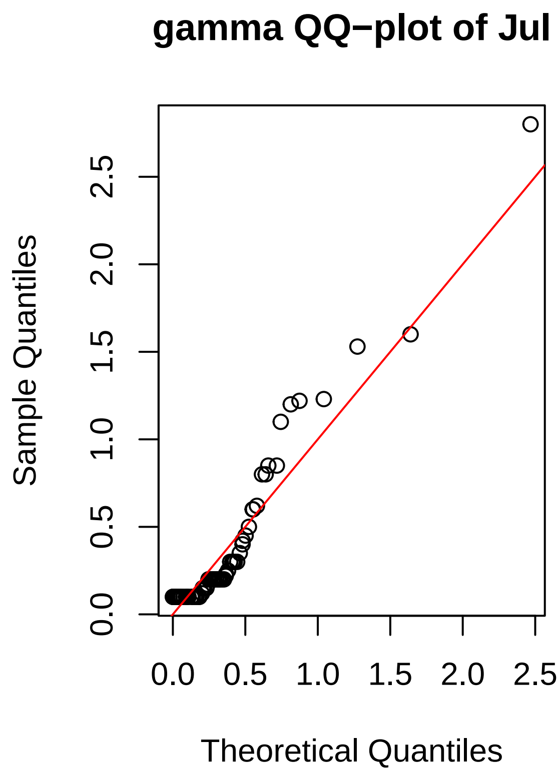


part d

Here we plot the gamma QQ-plot for both months and we can see that the gamma distribution is more adequate than the normal distribution for these two collections of data.

```
fit_Jan <- Jan.fit
fit_Jul <- July.fit
par(mfrow = c(1, 2))
mean_Jan <- fit_Jan$estimate[1]/fit_Jan$estimate[2]
var_Jan <- (fit_Jan$sd)^2
probabilities = (1:length(Jan))/(length(Jan)+1)
gamma.quantiles = qgamma(probabilities, shape = mean_Jan^2/var_Jan, scale = var_Jan/mean_Jan)
plot(sort(gamma.quantiles), sort(Jan), xlab = 'Theoretical Quantiles', ylab = 'Sample Quantiles', main =
abline(0,1, col = "red")
mean_Jul <- fit_Jul$estimate[1]/fit_Jul$estimate[2]
var_Jul <- (fit_Jul$sd)^2
probabilities = (1:length(Jul))/(length(Jul)+1)
gamma.quantiles = qgamma(probabilities, shape = mean_Jul^2/var_Jul, scale = var_Jul/mean_Jul)
plot(sort(gamma.quantiles), sort(Jul), xlab = 'Theoretical Quantiles', ylab = 'Sample Quantiles', main =
abline(0,1, col = "red")
```





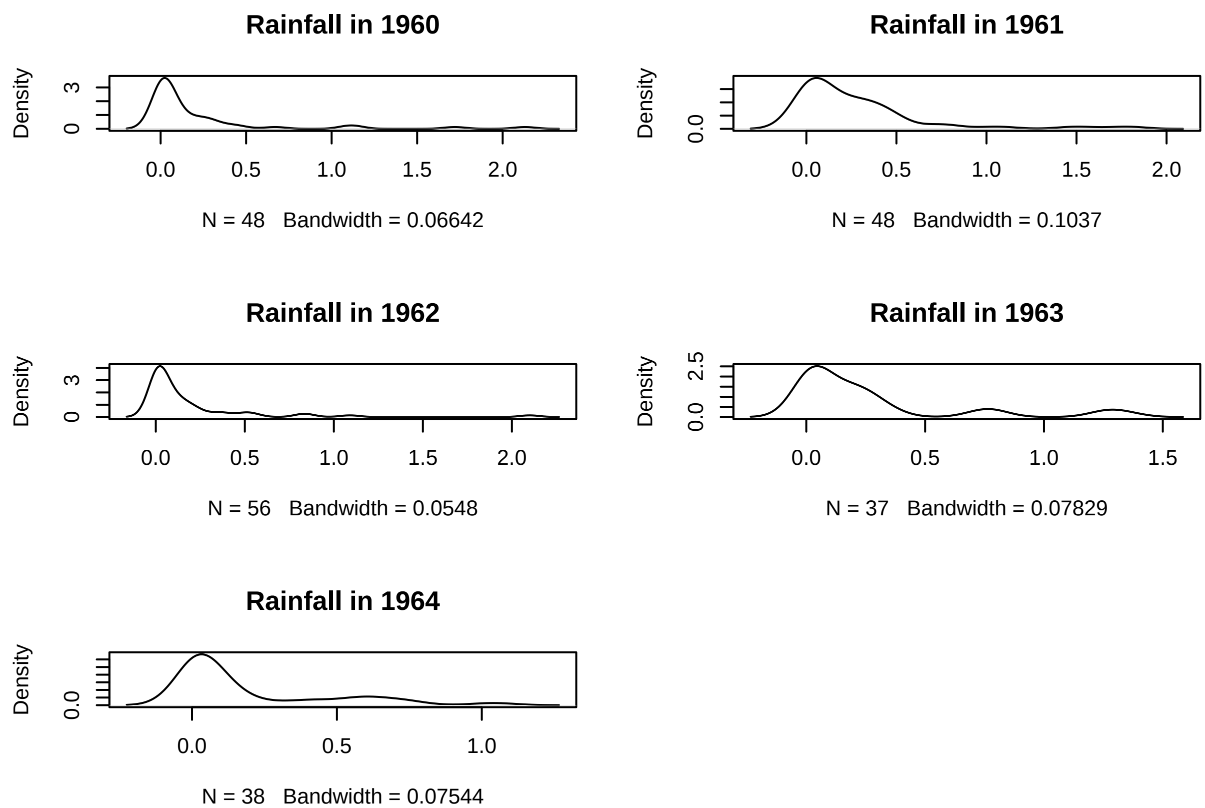
Illinois rainfall

part a

Questions: 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.

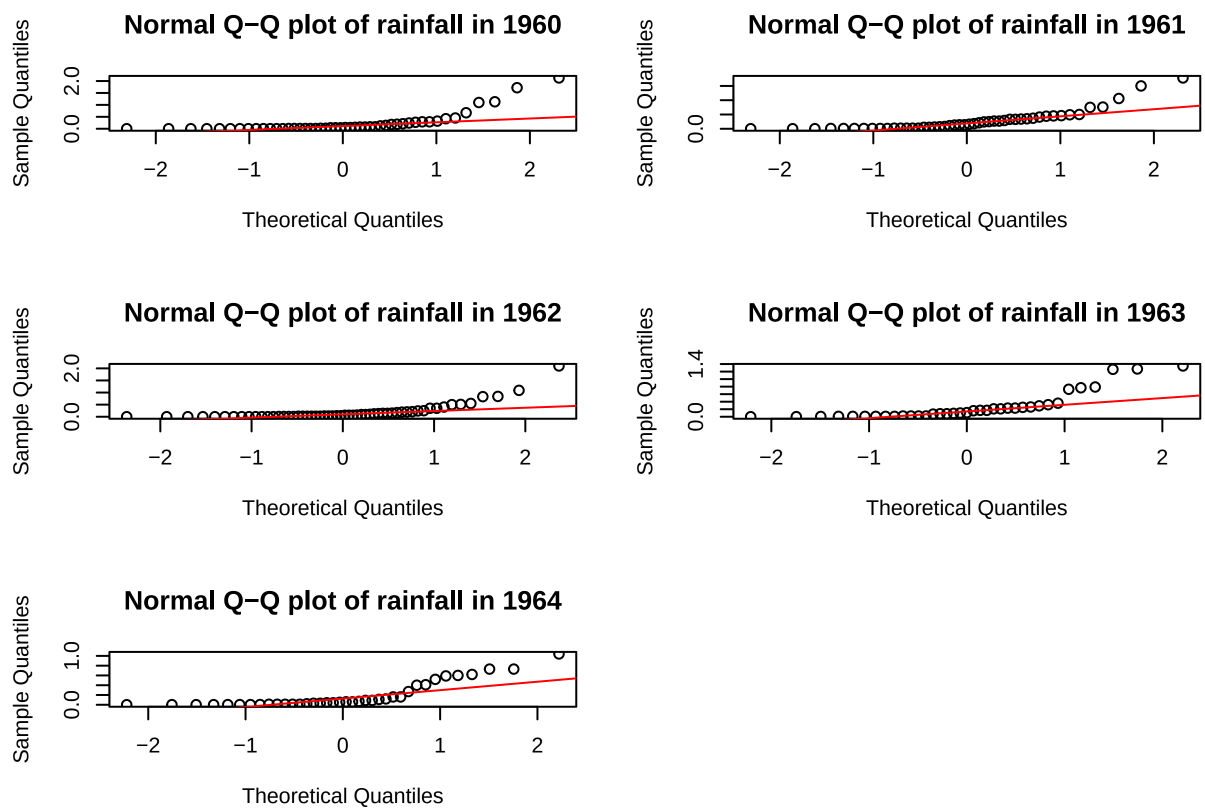
I firstly look into the distribution of the data using density plots.

```
rain <- read.xlsx('Illinois_rain_1960-1964.xlsx')
years <- c(1960, 1961, 1962, 1963, 1964)
par(mfrow = c(3, 2))
for (i in 1:5){
  plot(density(na.omit(rain[,i])), main = paste("Rainfall in", years[i]))
}
```



Then I use the Q Q plots to check whether they are normally distributed, and the graphs suggest that are not.

```
par(mfrow = c(3, 2))
for (i in 1:5) {
  temp <- na.omit(rain[,i])
  qqnorm(temp, pch = 1, main = paste("Normal Q-Q plot of rainfall in", years[i]))
  qqline(temp, col = "red", lwd = 1)
}
```



I notice that the data seems to have the same pattern as the data in the exercise 4.27, so I try to fit a gamma model using MLE. After checking the model result, we can see the log-likelihood and information about other parameters. I also show the confidence interval of my estimations, it seems that the confidence interval is quite narrow, which means there are only a few uncertainties. Then I look at the model checking graphs, they all suggest that the model fits well. Thus I can conclude that my model estimation is reliable.

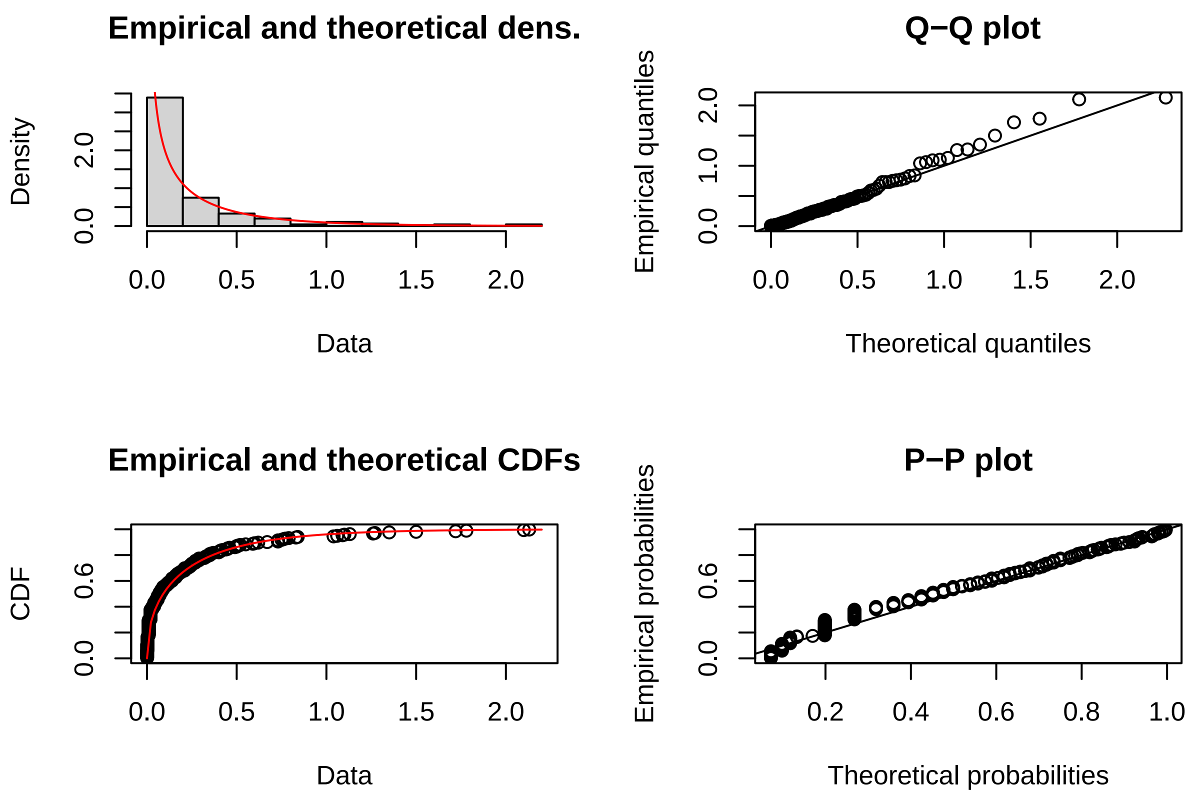
```
fit <- fitdist(c(na.omit(unlist(rain))), 'gamma', method='mle')
summary(fit)
```

```
## 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
```

```
summary(bootdist(fit))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median     2.5%     97.5%
## shape 0.4451446 0.383265 0.5196659
## rate  1.9932002 1.565640 2.6097674
```

```
plot(fit)
```



Part b

Questions: 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 identify the wet and dry years, we need a standard to classify them. I think using the mean rainfall amount in a year and the overall mean together would work. By reading the following table we can see that compared with the overall rainfall amount, the amount of 1962 and 1964 are noticeable less, so they are considered dry years. And 1961 and 1963 are considered wet years. Although 1962 has the most storms, it is a dry year. And 1963 only has 38 storms but it is a wet year. So, we can conclude that the number of storms does not affect the rainfall amount.

```
overall_mean <- fit$estimate[1] / fit$estimate[2]
each_mean <- apply(rain, 2, mean, na.rm = TRUE)
num_storm <- apply(!is.na(rain), 2, sum) %>% round(1)
result <- as.data.frame(t(data.frame(mean = c(each_mean, overall_mean),
                     storm = c(num_storm, mean(num_storm)))))
names(result)[6]='overall mean'
flextable(result)
```

1960	1961	1962	1963	1964	overall mean
0.2202917	0.2749375	0.18475	0.2624324	0.1871053	0.2243635
48.0000000	48.0000000	56.00000	37.0000000	38.0000000	45.4000000

Part c

Question: To what extent do you believe the results of your analysis are for generalization? 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.

Answer: Although it shows that our model is fitting well, it is only based on five years' data. So, if we really want to generalize our findings, we need more rainfall data from multiple years and multiple locations, and then rebuild our model. So, the next step would be to collect more data and update the model. In the article by Floyd Huff, it provides strong evidence that even a 100- year record of point rainfall may be misleading in estimating the frequency of extreme rainfall events. This shows that it is really hard to generalize the result and build a weather forecast system based only on partial data.

Reference:

1. https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r
2. https://www.r-bloggers.com/2015/11/profile-likelihood/
3. https://stackoverflow.com/questions/24211595/order-statistics-in-r?msclki d=fd6683dac56711ecbfcea9bd8a172395
4. Special thanks to my folk Yuli Jin who has provided many ideas and advice to this project. He helps me to get a deeper insight into the Gamma distribution and how it differs from the normal distribution. Meanwhile, I also have learned how to apply the gamma distribution when normal distribution does not work.