MA_677_final

2022-05-01

In All Likelihood

4.25

The density function for X_(n) is given by

$$f_{(k)}(x) = f_{(k)}(x) = nf(x)\binom{n}{k}F(x)^{k-1}(1-F(x))^{n-k} = \begin{cases} n\tbinom{n-1}{k-1}x^{k-1}(1-x)^{n-k} & \text{if 0 < x< 1} \\ \mathit{0} & \text{otherwise} \end{cases} \tag{1}$$

Thus, this density function is an example of beta distribution with r=k and s= n-k+1

when n=5, E represent the median gotten from beta distribution. medianUi represent the median from parameters.

```
E(2.5,5)
```

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

```
medianUi(2.5,5)
```

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

when n=10,E represent the median gotten from beta distribution. medianUi represent the median from parameters.

```
E(5,10)
```

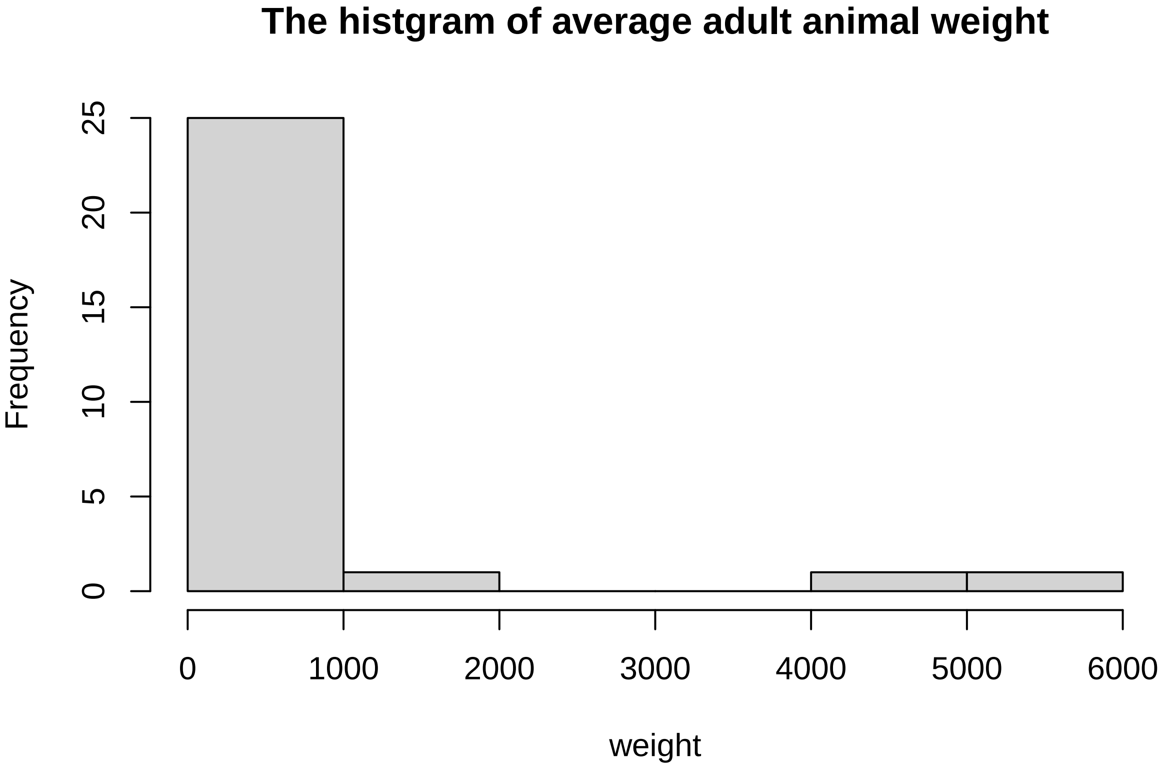
```
## [1] 0.4545455
```

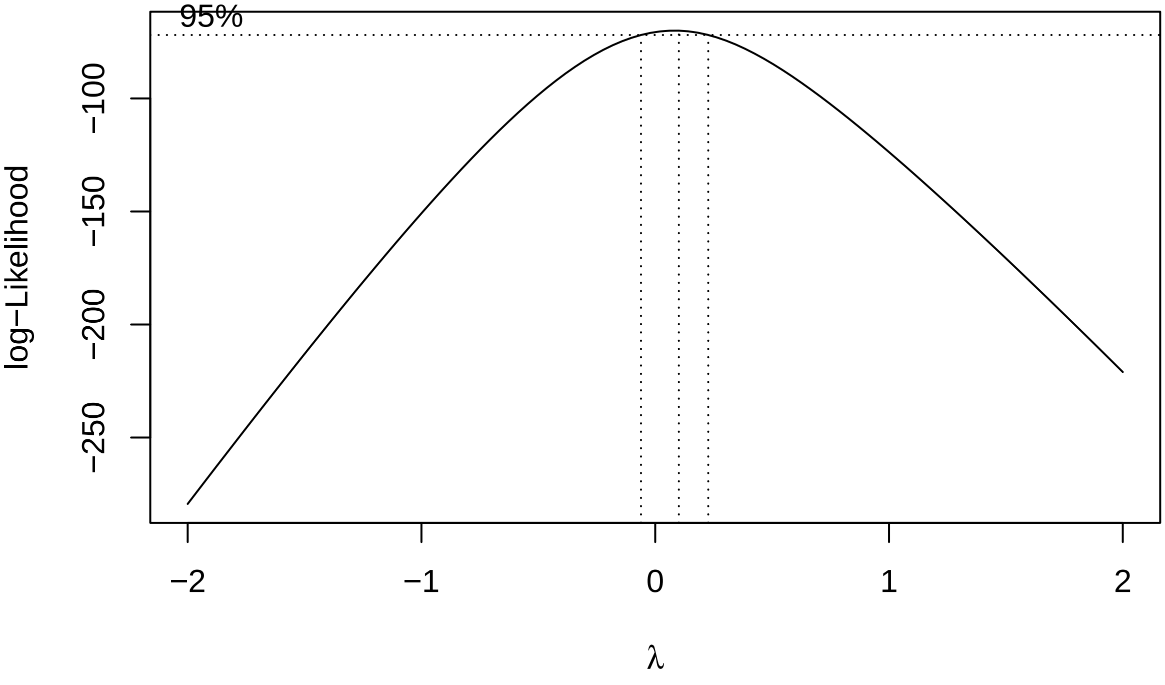
```
medianUi(5,10)
```

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

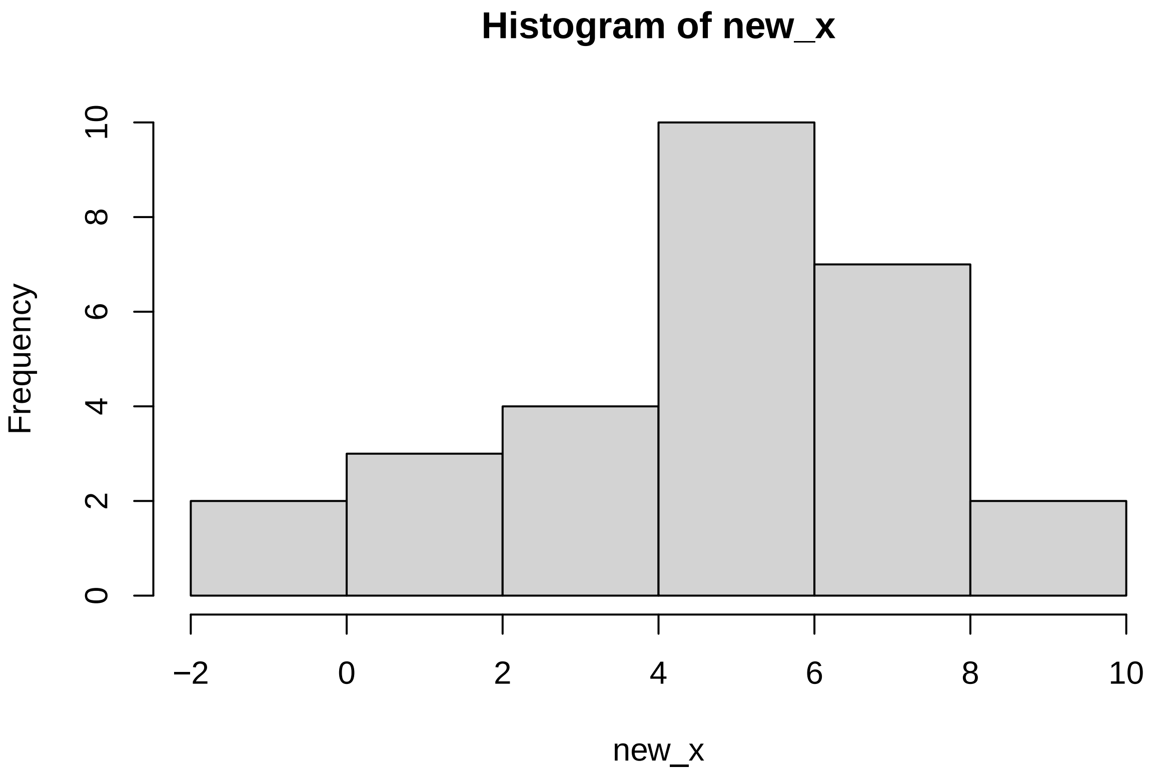
Thus, we can conclude that our approximation works well. Two sides of the equation are approximatly the same.

4.39





From above plot, we find that the distribution should be highly right skewness. Based on the outline of this question, we try to fit the Box-cox transformation. From below graph, the center dashed vertical line represents the estimated parameter λ and the others the 95% confidence interval of the estimation. The 0 is inside the confidence interval of the optimal λ and as the estimation of the parameter is really close to 0. Although our estimated λ is 0.1010101, our 95% confidence interval include 0. We still should use log transformation.



The last step is to based on shapiro test to test whether this data follow the normal distribution. The p-value is larger than the normal alpha value. Thus, we fail to reject the null hypothesis of normality. We can conclude that we have already transform our data approximately as normal distribution.

```
shapiro.test(new_x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_x
## W = 0.95787, p-value = 0.31
```

4.27

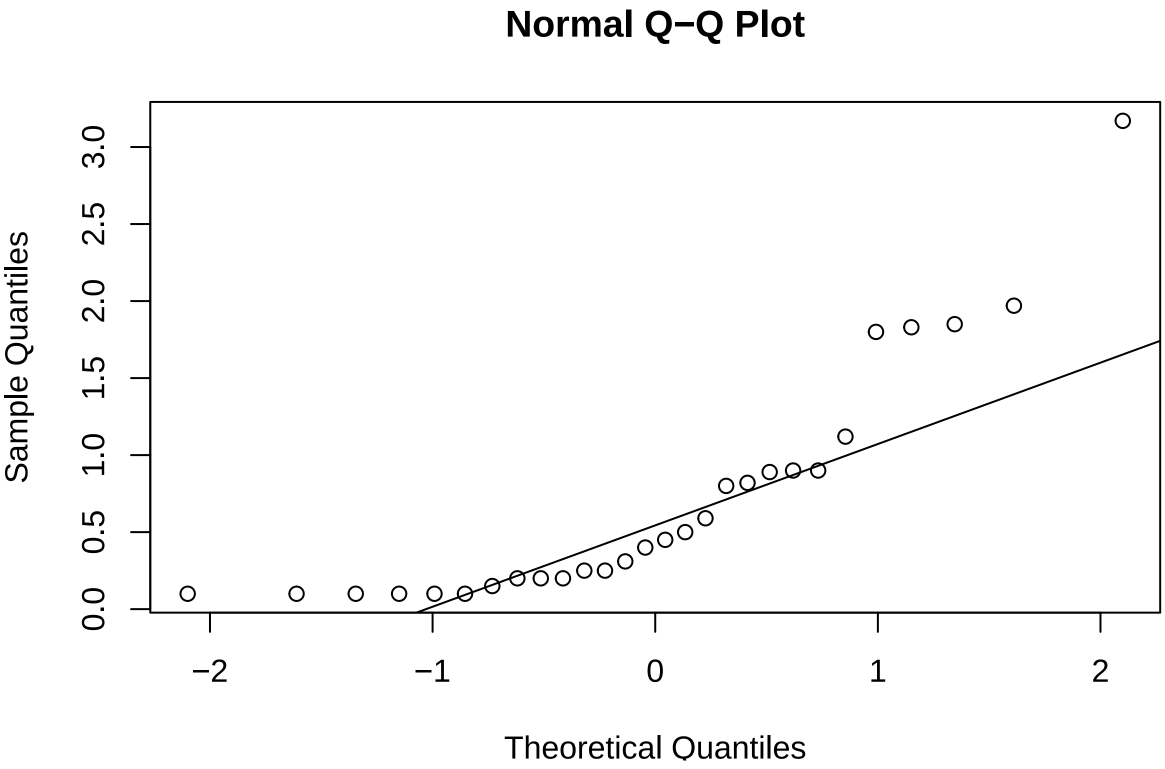
(a)

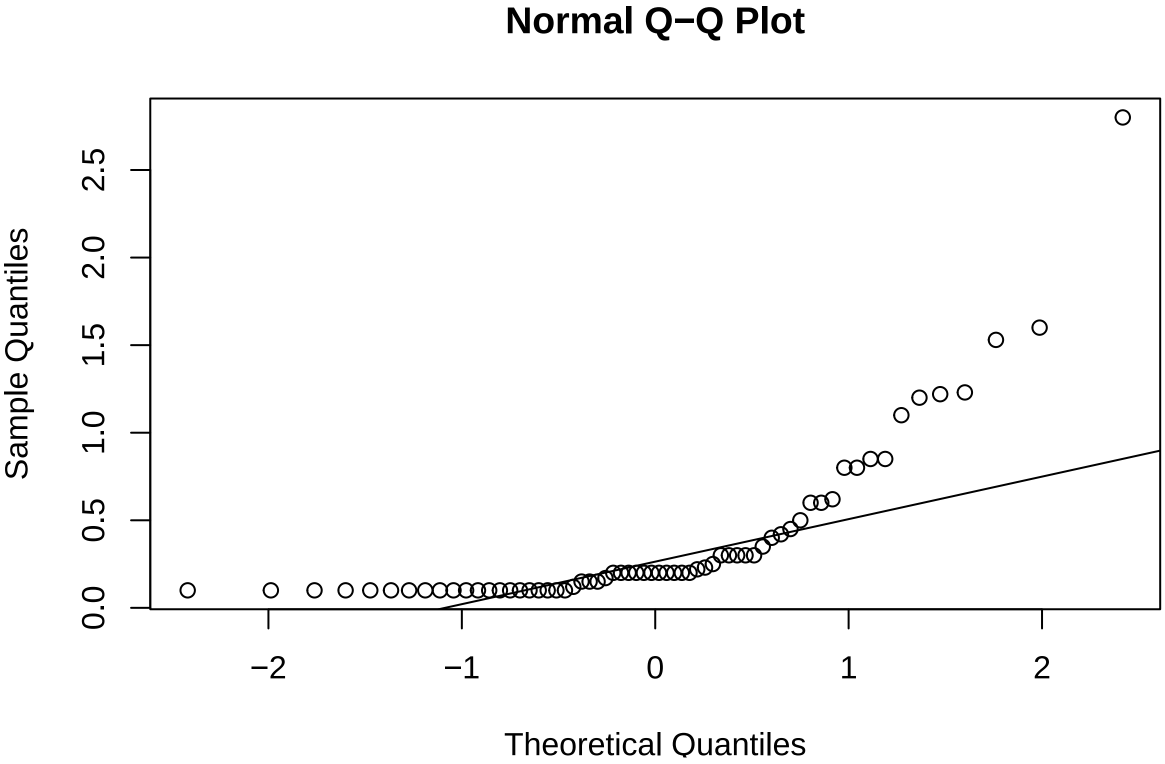
```
##       Jan              July
##  Min.   :0.1000   Min.   :0.1000
##  1st Qu.:0.1875   1st Qu.:0.1000
##  Median :0.4000   Median :0.2000
##  Mean   :0.7159   Mean   :0.3931
##  3rd Qu.:0.9000   3rd Qu.:0.4275
##  Max.   :3.1700   Max.   :2.8000
```

From above summary statistics, we find that the mean rainfall per storm in January is larger than that of in July.

(b)

After plotting the QQ plot of January and July, we find that both the rainfall in January and July both don't follow the normal distribution. We probably generalized models.





(c)

```
## [1] 1.056259 1.467754
```

```
## [1] 0.2498281 0.4397830
```

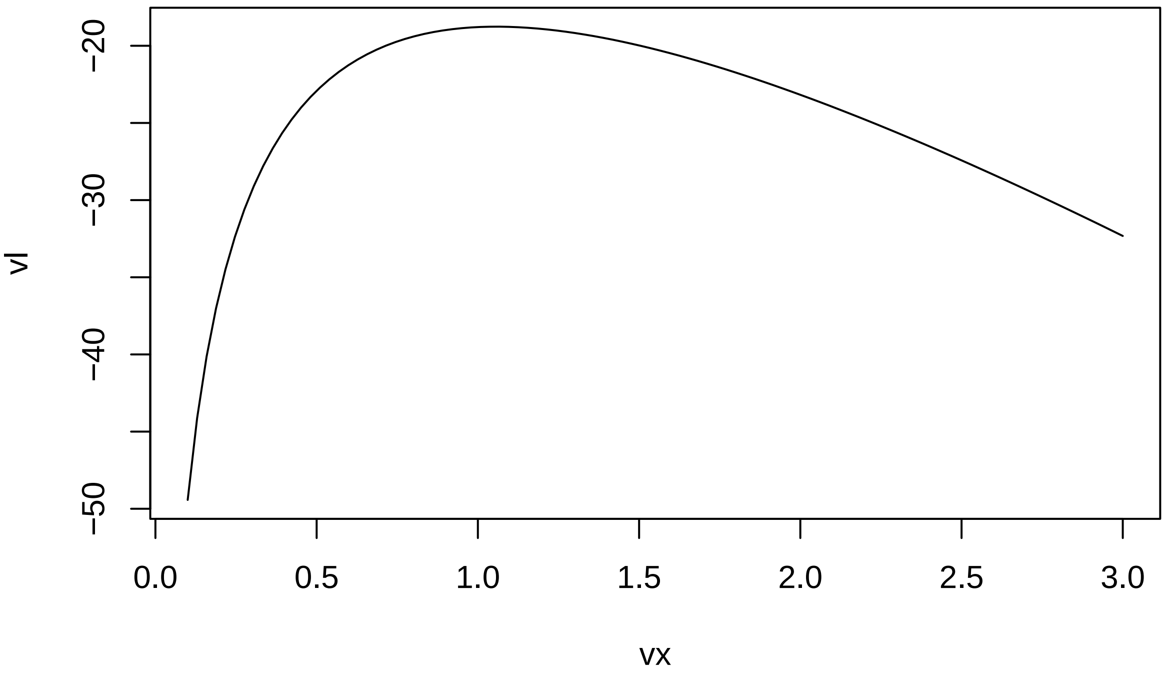
Thus, we estimate our two parameters for January is 1.0562594 and 1.4677541. The standard error is defined by the hessian matrix. The standard error of ML estimator is defined by the diagonal of the inverse of hessian matrix. The standard error is 0.2498281 and 0.439783.

Try to repeat the same step for July.

```
## [1] 1.196400 3.043306
```

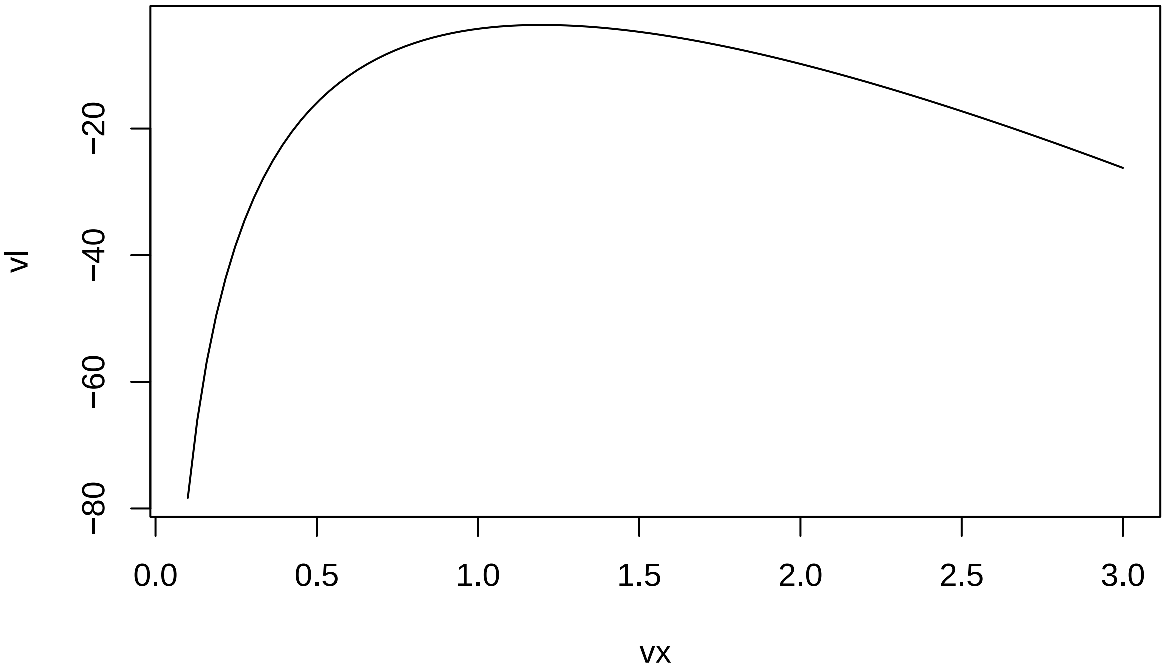
```
## [1] 0.1891734 0.5938089
```

Thus, we estimate our two parameters for July is 1.1964004 and 3.0433064. The standard error is defined by the hessian matrix. The standard error of ML estimator is defined by the diagonal of the inverse of hessian matrix. The standard error is 0.1891734 and 0.5938089.



```
## $par
## [1] 1.05625
##
## $value
## [1] 18.7616
##
## $counts
## function gradient
##       20       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Try to repeat the same step for July. Because July's likelihood is higher then that of January, July's model is bet-

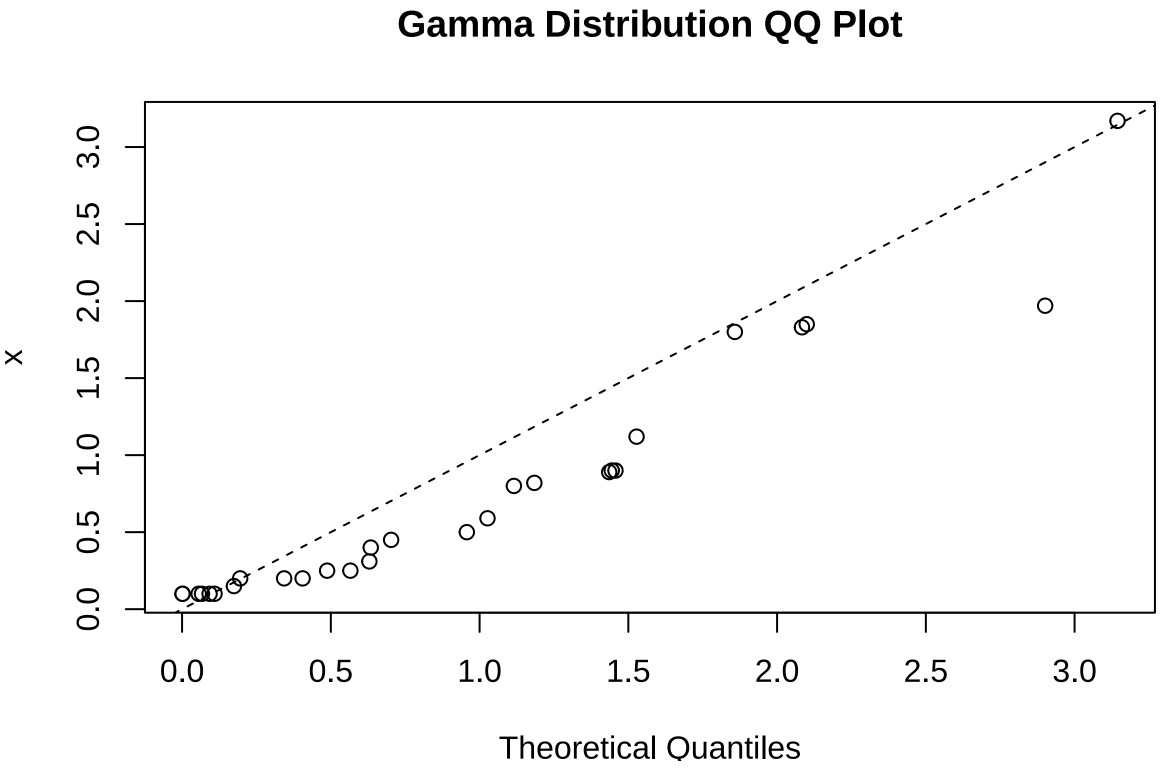


ter than January's.

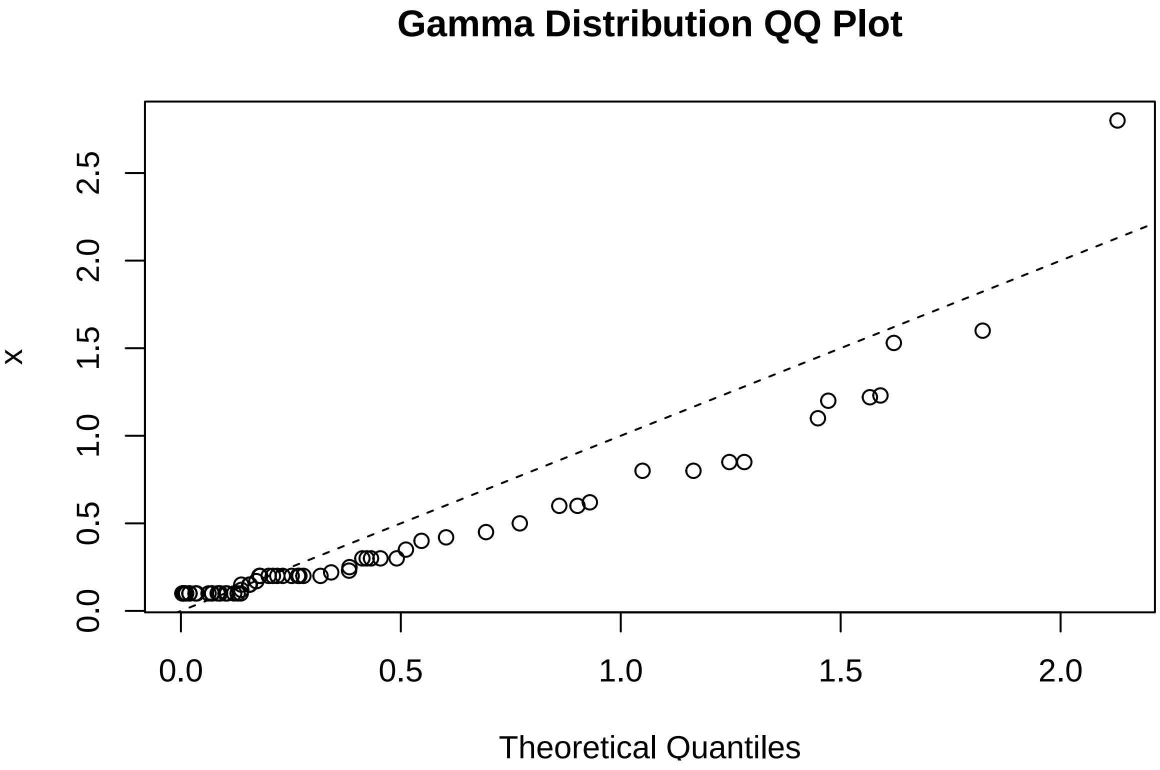
```
## $par
## [1] 1.196289
##
## $value
## [1] 3.634887
##
## $counts
## function gradient
##       22       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

(d)

Plot for January.



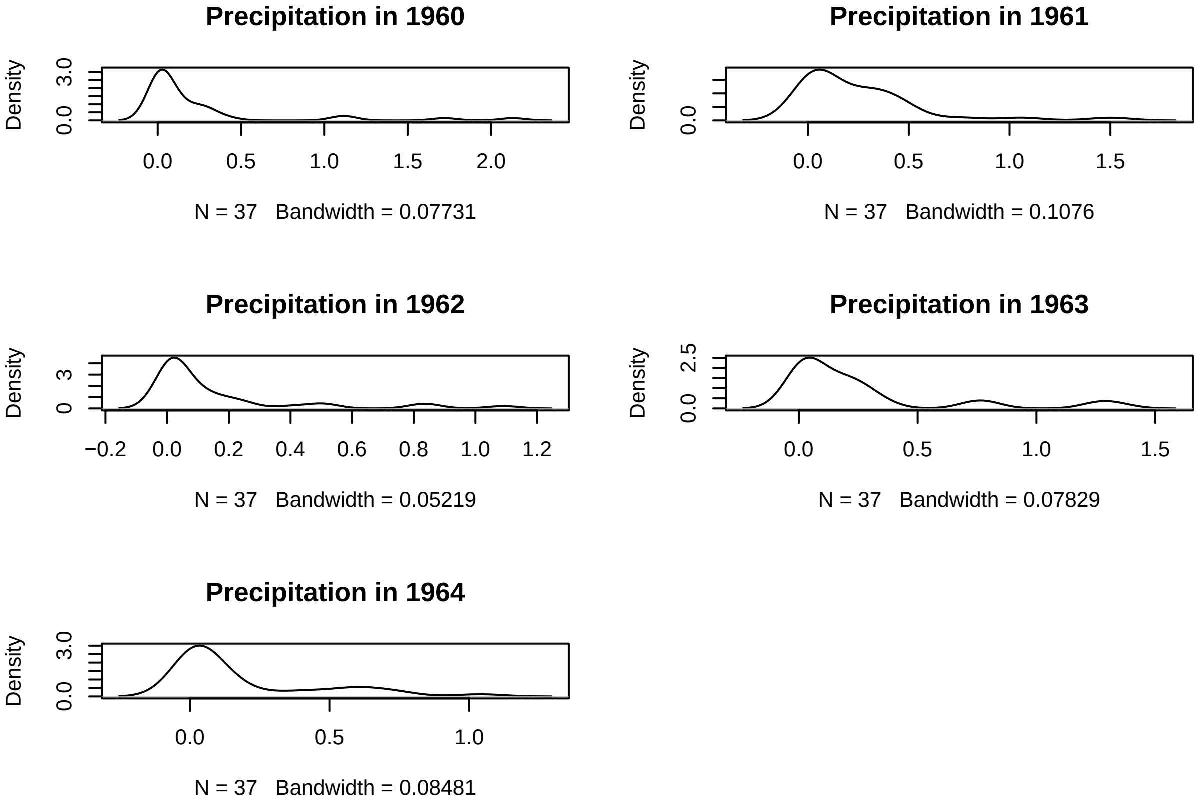
Plot for July.



From QQ plot, the Gamma model for July rely on more on the ab line, which means it have better adequacy of the gamma model.

Illinois rain

As the first step, we would like to visualization the rainfall data set. We find that most precipitation is right-skewed.



After plotting the density plot, we try to the distribution using MLE. The median and 95% confidence interval values are shown below. If we only based on the original variables without re sampling, our estimated parameters for shape and rate fall between the below 95% range.

```
## Loading required package: survival
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4524775 0.3856683 0.5345463
## rate  2.0332889 1.5393093 2.6620989
```

The definition of wet years and dry years come from Gao's analysis. "Years whose annual runoff was greater than the mean annual runoff over 1981–2010 were identified as wet years, while those that were less than the mean annual runoff were identified as dry years." (Gao et al., 2018). Using the previous naive model, we calculate each mean annual runoff for 1960 - 1964. Based on the below chart, we find that 1960,1962 and 1964 are dry years because the annual runoff is lower than the mean runoff. 1961 and 1963 are wet years because the annual runoff is larger than the mean runoff.

```
##      1960      1961      1962      1963      1964      mean
## 0.2202917 0.2749375 0.1847500 0.2624324 0.1871053 0.2233725
```

Based on the analysis of Huff (Changnon & Huff, 1967), "During this 5-year period, 34% of the total rainfall occurred in 5% of the storms, over 50% fell in 10% of the storms, and 75% occurred in 20% of the storms. Thus, it is apparent that relatively few storms established the air mass rainfall pattern during a rather substantial sampling period of 5 years, frequently considered adequate to evaluate seeding experiments" Also, combined with the idea illustrated in (Huff, 1994) that even 100 years record will be influenced by one extreme event. We agree with the idea that the wet year wet because individual storms produced more rain.

I think the results of my analysis is only limited to 1960-1964. Because we only simulate a very naive gamma model without re sampling and specify the distribution. The generalization for this analysis still need more steps to verify. The next step is to re sample for predicting. Also, figuring out each rainfall's time and other background information should be helpful. Moreover, based on Huff's analysis, the single rainfall for individual spot is largely different than others. Single spots cumulative rainfall data should useful to our analysis.

What I learned and future plan

Previously, I never think problems from the perspective of likelihood. To be more honest, I never care about the possible parameters before. I try to read the chapter 4 and revisit the process of models, which help to understand and identify the problem from the back side of the model. When I face problems, I try to revisit the definition of model and try to think why those model build on those parameter. Is there any assumption about the model? Such kind of information like that.

The next step for this project is already discussed in the previous section. The next step for me is to try to understand the models from another perspective.

Reference:

1. Gao, X., Chen, X., Biggs, T., & Yao, H. (2018). Separating wet and dry years to improve calibration of Swat in Barrett watershed, Southern California. Water, 10(3), 274. https://doi.org/10.3390/w10030274

2. Changnon, S. A., & Huff, F. A. (1967). The effect of natural rainfall variability in verification of rain modification experiments. Weather Modification Experiments, 177–198. https://doi.org/10.1525/9780520313903-013

3. Huff, F. A. (1994). Record-breaking Microstorm system supports new rainfall frequency estimates in Illinois. Bulletin of the American Meteorological Society, 75(7), 1223–1226. https://doi.org/10.1175/1520-0477(1994)075%3C1223:rbmssn%3E2.0.co;2

4. Huff, F. A. (1967). Time distribution of rainfall in heavy storms. Water Resources Research, 3(4), 1007–1019. https://doi.org/10.1029/wr003i004p01007

Appendix

```r
# The uniform pdf
f <- function(x, mu = 0, sigma = 1) dunif(x, mu, sigma)
# The uniform CDF
F <- function(x, mu = 0, sigma = 1) punif(x, mu, sigma, lower.tail = FALSE)
# The density function for X_(n)
f_k <- function(x,r,n) {
  x*f(x) *(1-F(x))^(r - 1) * (F(x))^(n - r)
}

# The median of beta function
E <- function(r, n){
(1/beta(r, n - r + 1)) * integrate(f_k, -Inf, Inf, r, n)$value
}

# The estimated of median Ui.
medianUi <- function(k, n){
m <- (k - 1/3) / (n + 1/3)
return(m)
}
E(2.5,5)
```

```r
medianUi(2.5,5)
E(5,10)
medianUi(5,10)
# Import the data set of 4.39
library(MASS)
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
hist(weight,main = "The histgram of average adult animal weight")
# box cox test
bc<- boxcox(weight~ 1)
lambda <- bc$x[which.max(bc$y)]
# histogram after transformation
new_x <- log(weight)
hist(new_x)
shapiro.test(new_x)
# Import the data of 4.27
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)
d <- data.frame(cbind(Jan,July))
summary(d)
#QQ plot for January
qqnorm(Jan)
qqline(Jan)
#QQ plot for July
qqnorm(July)
qqline(July)
# build an approximated confidence interval for alpha.
log_lik <- function(x){
alpha <- x[1]
beta <- x[2]
logL <- dgamma(month, shape = alpha, scale = 1 / beta)
result <- -1 * sum(log(logL))
return(result)
}
# We guess the initial point is around 0.5.
p<- array(c(0.5, 0.5), dim = c(2, 1))
month <- Jan
ans_jan <- nlm(f = log_lik, p, hessian = T)
ans_jan$estimate
sqrt(diag(solve(ans_jan$hessian)))
# We guess the initial point is around 0.4.
p<- array(c(0.5, 0.5), dim = c(2, 1))
month <- July
ans_jan <- nlm(f = log_lik, p, hessian = T)
ans_jan$estimate
sqrt(diag(solve(ans_jan$hessian)))
```

```r
# plot for likehood for January
prof_log_lik <- function(y){
a <- (optim(1, function(z) - sum(log(dgamma(x, y, z)))))$par
result <- -sum(log(dgamma(x, y, a)))
return(result)
}
x <- Jan
vx=seq(0.1,3,length=100)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l")
optim(1,prof_log_lik)
# plot for likehood for July
x <- July
vx=seq(0.1,3,length=100)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l")
optim(1,prof_log_lik)
qqGamma <- function(x)
{
aa = (mean(x))^2/ var(x)
ss = var(x) / mean(x)
test = rgamma(length(x), shape = aa, scale = ss)
qqplot(test, x, xlab = "Theoretical Quantiles",
main = "Gamma Distribution QQ Plot")
abline(0,1, lty = 2)
}
qqGamma(Jan)
qqGamma(July)
library("openxlsx")
rain <- read.xlsx("~/Downloads/Illinois_rain_1960-1964(2).xlsx")
par(mfrow = c(3, 2))
plot(density(na.omit(rain)$`1960`), main = "Precipitation in 1960")
plot(density(na.omit(rain)$`1961`), main = "Precipitation in 1961")
plot(density(na.omit(rain)$`1962`), main = "Precipitation in 1962")
plot(density(na.omit(rain)$`1963`), main = "Precipitation in 1963")
plot(density(na.omit(rain)$`1964`), main = "Precipitation in 1964")
library(fitdistrplus)
fit_rain <- fitdist(unlist(na.omit(rain)), "gamma", method = "mle")
summary(bootdist(fit_rain))
mean <- fit_rain$estimate[1] / fit_rain$estimate[2]
mean_df <- apply(rain, 2, mean, na.rm = TRUE)
mean_df <- c(mean_df, as.numeric(mean))
names(mean_df) <- c("1960", "1961", "1962", "1963", "1964", "mean")
mean_df
```