MA677 Final Project

Yujia Wang

5/12/2022

Abstract

I completed all the "In All Likelihood" topics, as well as the insurance claims and species discovery of "Introduction to Empirical Bayesian".

Exercise

Exercise 4.25 (P112)

```r
# pdf of standard uniform distribution
f <- function(x, a=0, b=1){
  dunif(x, a, b)
}

# cdf of standard uniform distribution
F <- function(x, a=0, b=1){
  punif(x, a, b, lower.tail=FALSE)
}

# Distribution of order statistics
# Based on Exercise 2.4 P48
integrand <- function(x,r,n) {
  x * (1 - F(x))^(r-1) * F(x)^(n-r) * f(x)
}

# Expected value of sample size = n & the rth largest order statistic
expectation <- function(r,n) {
  (1/beta(r, n-r+1)) * integrate(integrand, lower=-Inf, upper=Inf, r, n)$value
}
# ( InF = infinity )
# ( expectation = median{U(i)} )

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

```
expectation(2.5,5) # when n = 5
```

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

```
approximation(2.5,5)
```

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

```
expectation(5,10) # when n = 10
```

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

```
approximation(5,10)
```

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

The result shows that expectation = approximation when n is different value, which means “Expected value of order statistics = median approximation” is valid.

Exercise 4.27 (P112)

```
# input data
january <- 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,
july <- c(0.30,0.22,0.10,0.12,0.20,0.10,0.10,0.10,0.10,0.10,0.10,0.17,0.20,2.80,0.85,0.10,0.10,1.23,0.4
```

(a)

```
summary(january)
```

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

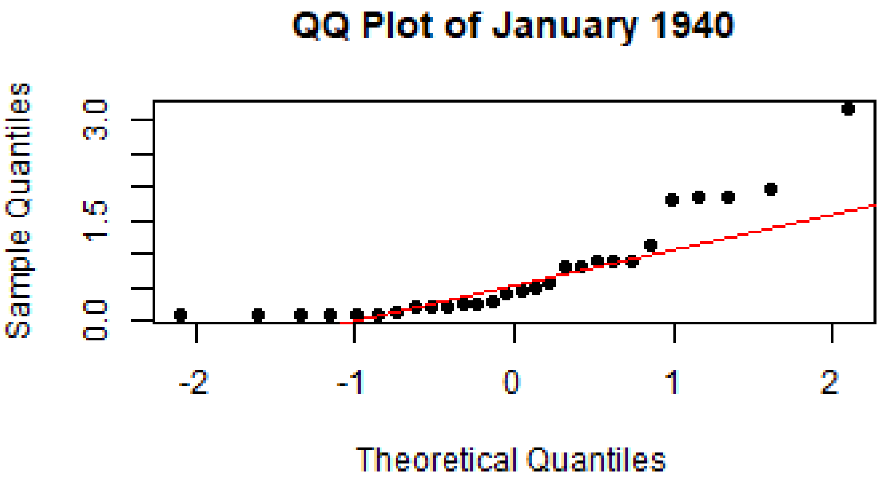
```
summary(july)
```

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

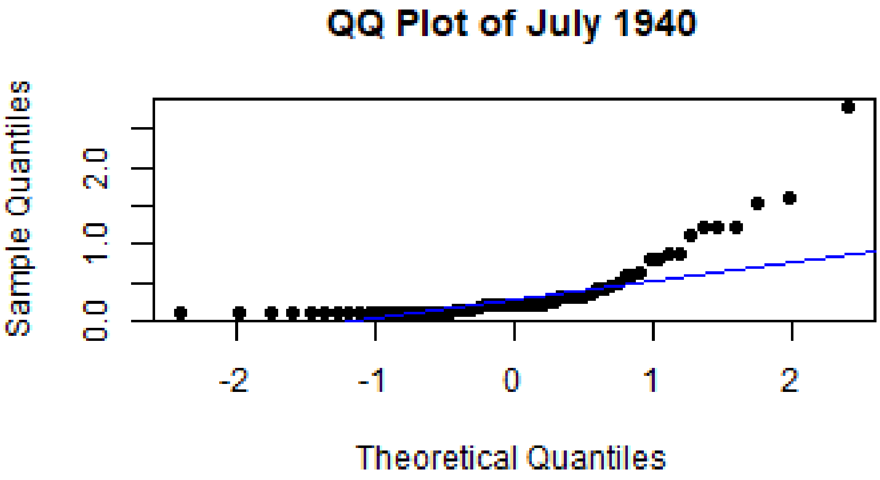
The average amount of rainfall in January 1940 is higher than that in July 1940.

(b)

```
qqnorm(january, main = "QQ Plot of January 1940",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles", pch = 16)
qqline(january, col = "red", lwd = 1)
```



```
qqnorm(july, main = "QQ Plot of July 1940",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles", pch = 16)
qqline(july, col = "blue", lwd = 1)
```



The QQ plots above are skewed to the right, and have long tails that have more extreme values in the largest parts than would be expected if they truly came from a Normal distribution (the smallest parts also have larger values than normal). So there are not uniform data. (P92) Therefore, as a nonnormal model, the gamma model is useful for positive outcome data. (P93)

(c)

```
fit1 <- fitdist(january,'gamma','mle')
summary(fit1)
```

```
## 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
```

```
fit2 <- fitdist(july,'gamma','mle')
summary(fit2)
```

```
## 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
```

AIC and BIC: the model in July is better than January.

```
exp(fit1$loglik)
```

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

```
exp(fit2$loglik)
```

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

MLE: July > January, so the model in July is better than January as well.

```
fit1$sd # standard errors
```

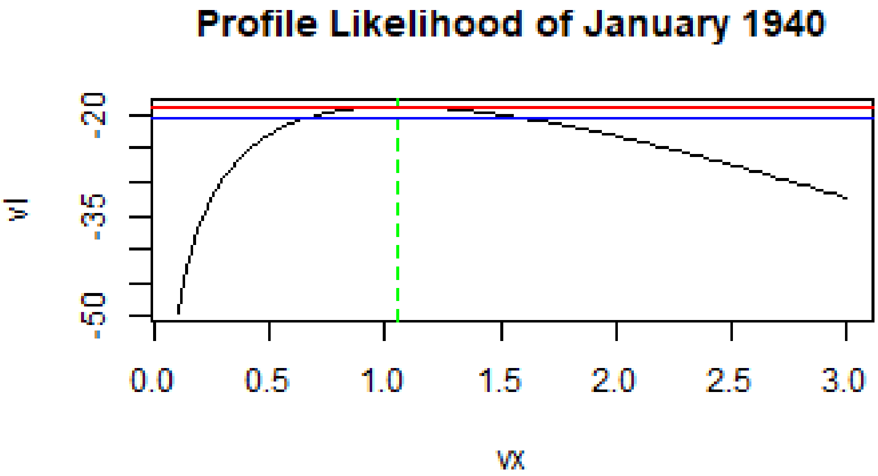
```
##     shape      rate
## 0.2497495 0.4396202
```

```
fit2$sd
```

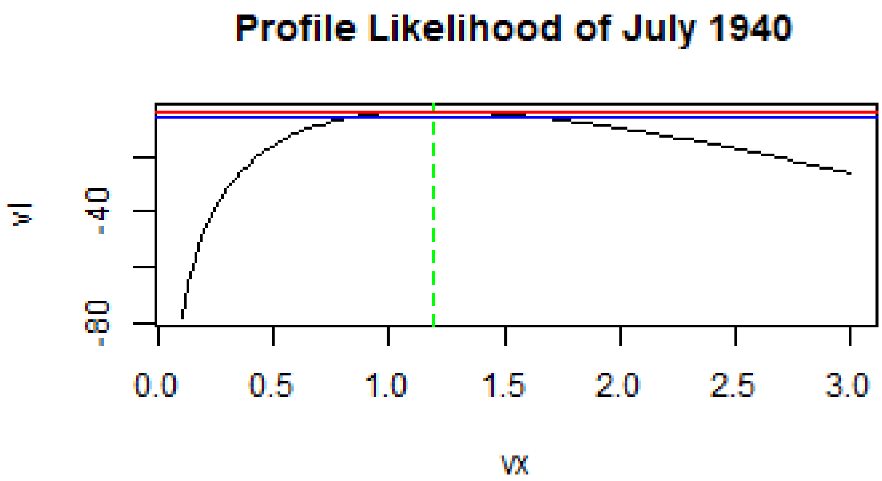
```
##     shape      rate
## 0.1891196 0.5936302
```

Standard error: alpha in July < January, beta in July > January.

```
# profile likelihoods for the mean parameters
prof_log_lik1 <- function(a){
  b = (optim(1, function(z) -sum(log(dgamma(january, a, z)))))$par
  return(-sum(log(dgamma(january, a, b))))
}
vx <- seq(.1, 3, length = 101)
vl <- -Vectorize(prof_log_lik1)(vx)
plot(vx, vl, type = "l", main = "Profile Likelihood of January 1940")
abline(v = optim(1, prof_log_lik1)$par,lty = 2, col = "green")
abline(h = -optim(1, prof_log_lik1)$value, col = "red")
abline(h = -optim(1, prof_log_lik1)$value-qchisq(.95, 1)/2, col = "blue")
```

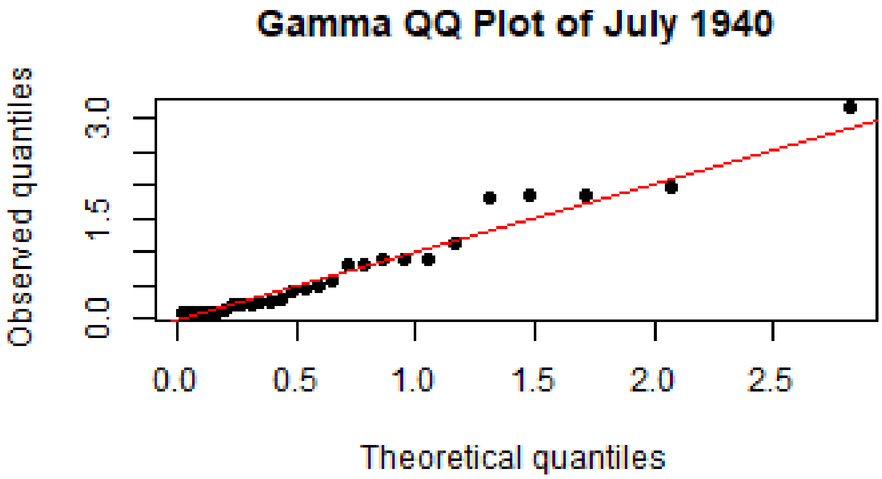


```
prof_log_lik2 <- function(a){
  b = (optim(1, function(z) -sum(log(dgamma(july, a, z)))))$par
  return(-sum(log(dgamma(july, a, b))))
}
vx <- seq(.1, 3, length = 101)
vl <- -Vectorize(prof_log_lik2)(vx)
plot(vx, vl, type = "l", main = "Profile Likelihood of July 1940")
abline(v = optim(1, prof_log_lik2)$par,lty = 2, col = "green")
abline(h = -optim(1, prof_log_lik2)$value, col = "red")
abline(h = -optim(1, prof_log_lik2)$value-qchisq(.95, 1)/2, col = "blue")
```

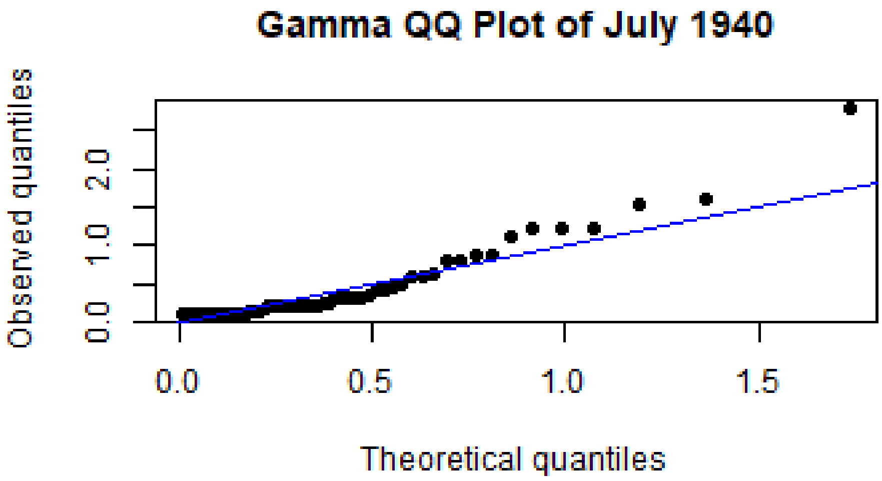


(d)

```
x <- sort(january)
x0 <- qgamma(ppoints(length(x)),
             shape = fit1$estimate[1],
             rate = fit1$estimate[2]);
plot(x = x0, y = x, main = "Gamma QQ Plot of July 1940",
     xlab = "Theoretical quantiles", ylab = "Observed quantiles", pch = 16)
abline(a = 0, b = 1, col = "red", lwd = 1)
```



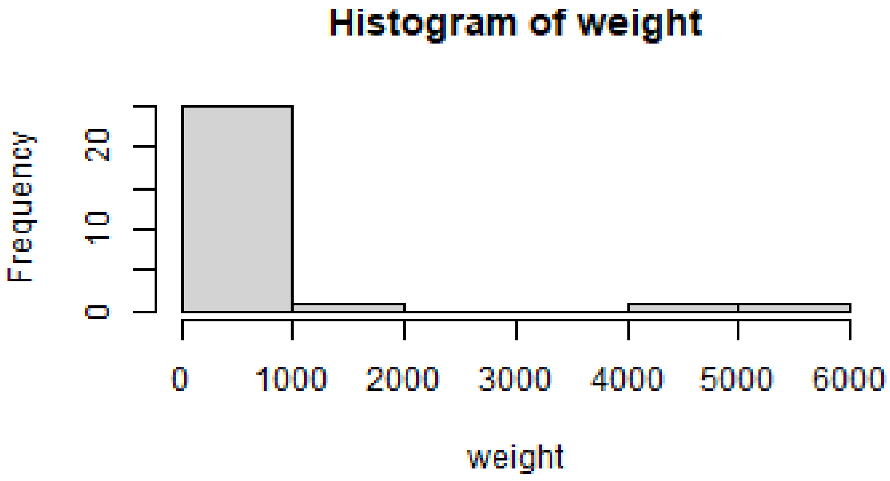
```
x <- sort(july)
x0 <- qgamma(ppoints(length(x)),
             shape = fit2$estimate[1],
                rate = fit2$estimate[2]);
plot(x = x0, y = x, main = "Gamma QQ Plot of July 1940",
     xlab = "Theoretical quantiles", ylab = "Observed quantiles", pch = 16)
abline(a = 0, b = 1, col = "blue", lwd = 1)
```



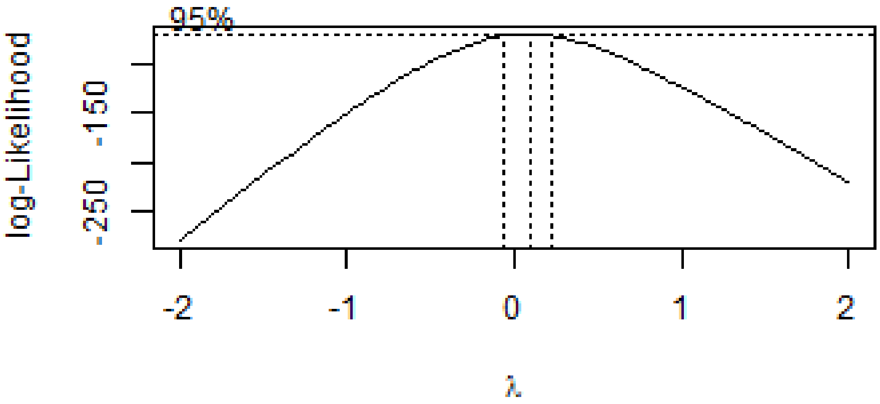
Using a gamma model is valid, espacially for July data.

Exercise 4.39 (P114)

```
# input 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,
hist(weight)
```



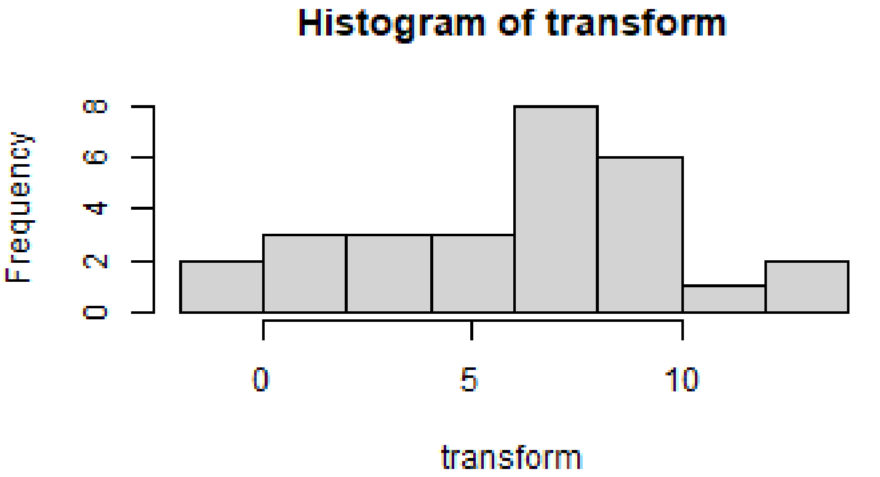
```
# Box-cox transform
bct <- boxcox(lm(weight ~ 1))
```



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

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

```
# transform x to new x
transform <- (weight ^ lambda - 1) / lambda
hist(transform)
```

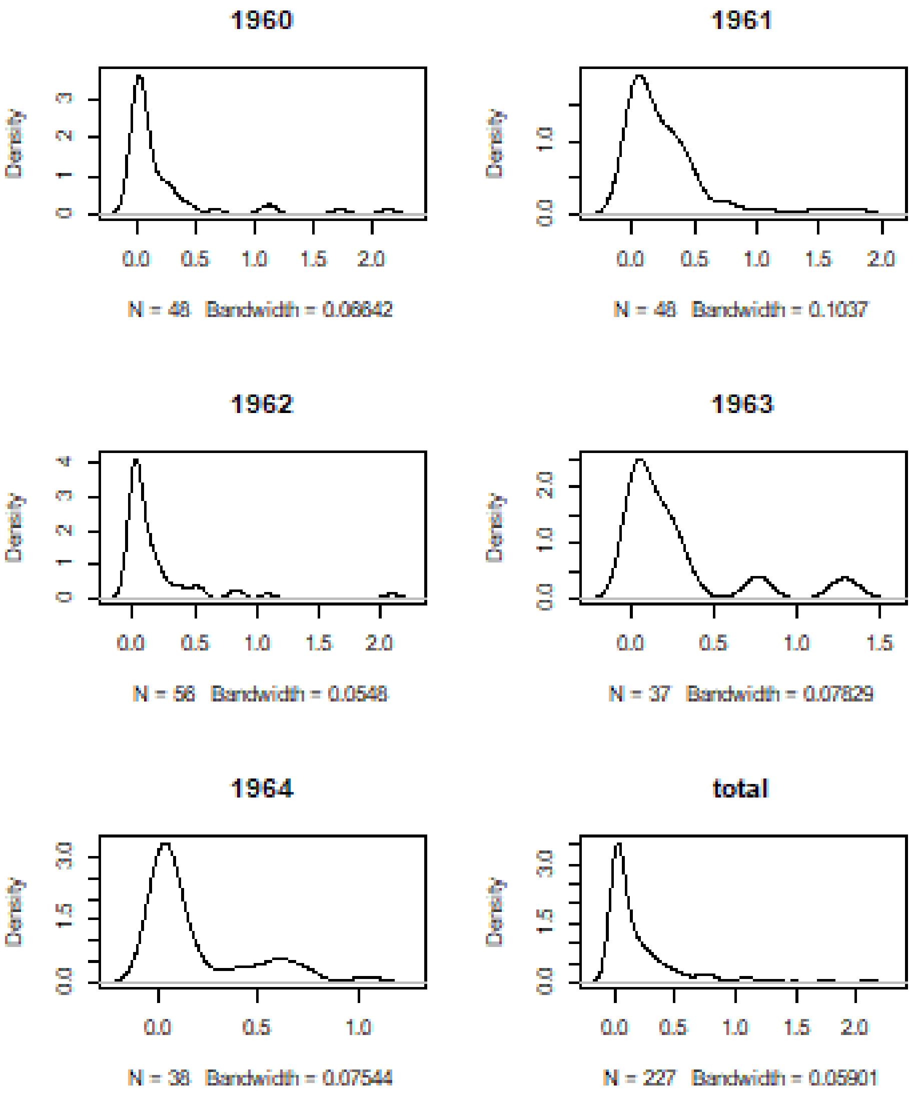


After the box-cox transformation, the transformation is valid because the 95% confidence interval does not include 1. And after the next transformation, the data closes to normal distribution.

In All Likelihood

Section 1

```
#setwd("E:/BU Study_2022/677_Final Project")
rain <- read_xlsx('Illinois_rain_1960-1964.xlsx')

# density plots for each year and total data
par(mfrow = c(3,2))
plot(density(rain$`1960`%>%na.omit()),main = "1960")
plot(density(rain$`1961`%>%na.omit()),main = "1961")
plot(density(rain$`1962`%>%na.omit()),main = "1962")
plot(density(rain$`1963`%>%na.omit()),main = "1963")
plot(density(rain$`1964`%>%na.omit()),main = "1964")
plot(density(unlist(rain)%>%na.omit()),main = "total")
```



Gamma distribution is suitable because of the distribution density shape.

```
rain2 <- unlist(rain) %>% na.omit()
rain2 <- as.numeric(rain2)

fit3 <- fitdist(rain2,distr = "gamma",method = "mle") #MLE estimation
summary(bootdist(fit3)) # bootdist: get confidence interval of parameters
```

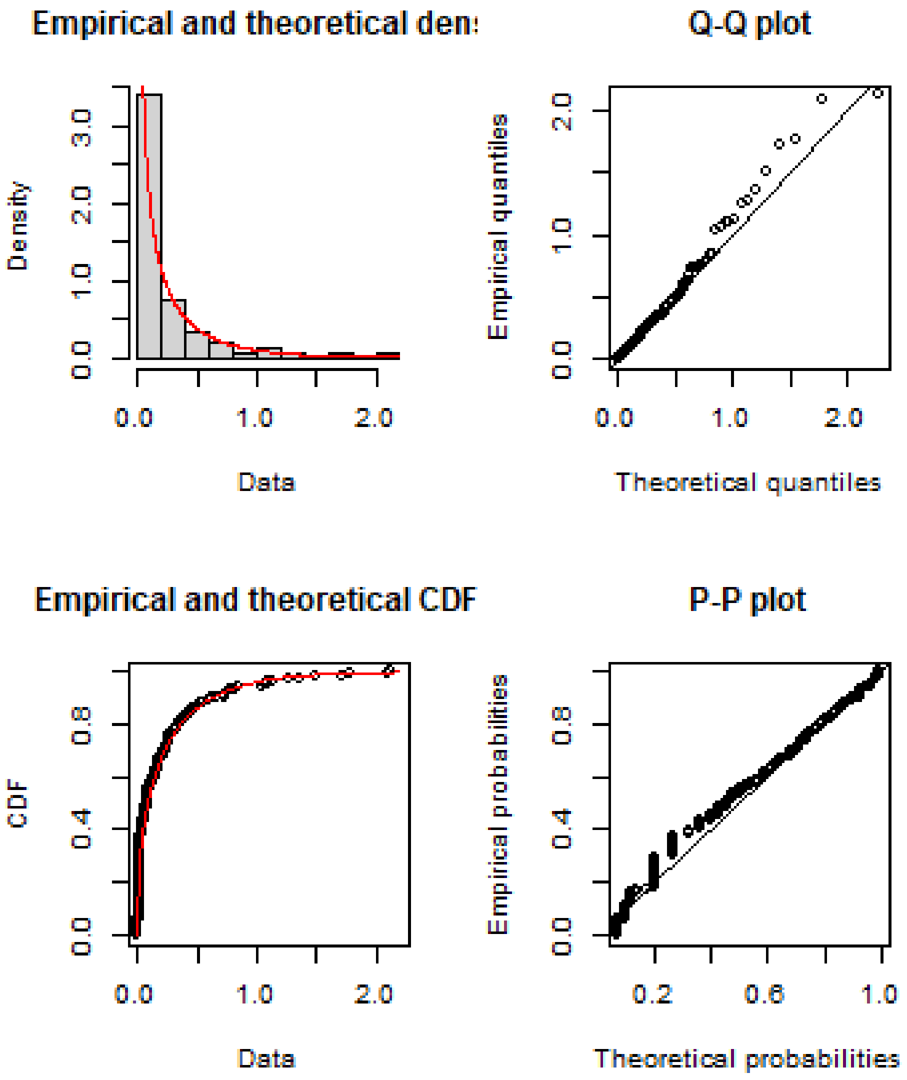
```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4439173 0.3841754 0.5196698
## rate  1.9868252 1.5384780 2.5214159
```

```
fit4 <- fitdist(rain2,distr = "gamma",method = "mse") #MSE estimation
summary(bootdist(fit4))
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median     2.5%     97.5%
## shape 0.7164466 0.614732 0.8452059
## rate  1.3454173 1.084707 1.6478014
```

MLE: the 95% confidence interval from bootstrap sample is (0.3834941, 0.5209687) for alpha, (1.5549701, 2.5343410) for beta. Parametric bootstrap median is 0.4432958 for alpha, 1.9748966 for beta. MSE: the 95% confidence interval from bootstrap sample is (0.6182149, 0.8404227) for alpha, (1.0536660, 1.6636052) for beta. Parametric bootstrap median is 0.7164441 for alpha, 1.3362024 for beta. So the MLE estimates have narrow interval which means lower variances and the confidence interval indicates that the estimation is valid. In conclusion, fit model using gamma distribution with MLE is suitable.

```
plot(fit3)
```



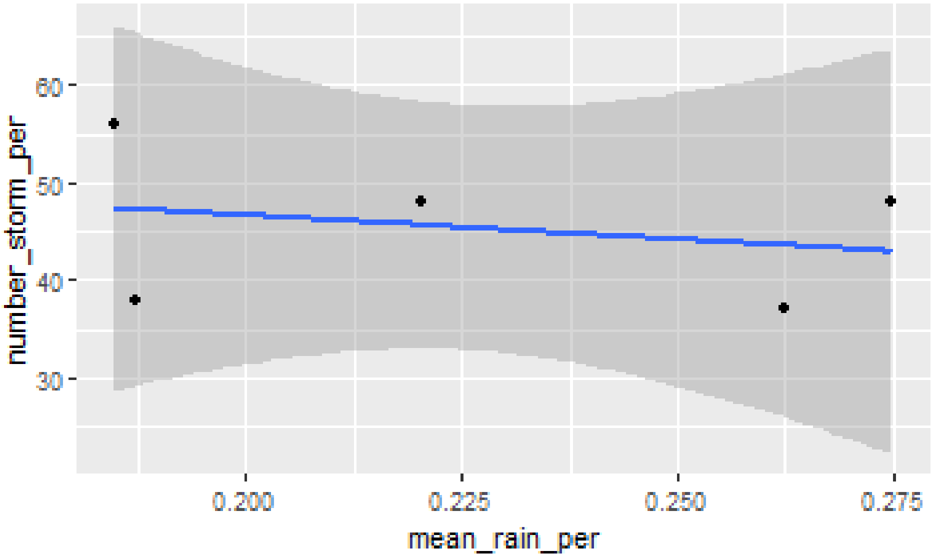
Section 2

```
# average rainfall of 5 years
mean_rain <- fit3$estimate[1]/fit3$estimate[2] # mu = alpha / beta

# average rainfall per year
mean_rain_per = c(apply(rain,2,mean,na.rm=TRUE),mean_rain)%>%round(4)
names(mean_rain_per) <- c('1960', '1961', '1962', '1963', '1964', 'summary')

# storms number of 5 years
number_storm <- length(rain2)/5

# storms number of each year
number_storm_per <- c(apply(!is.na(rain), 2, sum, na.rm = TRUE), number_storm)
names(number_storm_per) <- c('1960', '1961', '1962', '1963', '1964', 'summary')
```

```
# output
df <- rbind(mean_rain_per, number_storm_per)
df <- t(df)
df <- as.data.frame(df[-6,])
knitr::kable(rbind(mean_rain_per, number_storm_per))
```

	1960	1961	1962	1963	1964	summary
mean_rain_per	0.2203	0.2749	0.1848	0.2624	0.1871	0.2244
number_storm_per	48.0000	48.0000	56.0000	37.0000	38.0000	45.4000

```
ggplot(data=df, aes(x=mean_rain_per, y=number_storm_per)) +
  geom_smooth(method='lm', formula= y~x)+
  geom_point()
```



1962 and 1964 were below average and dry years, 1961 and 1963 were above average and wet years, and 1960 was near average. In 1961, the highest rainfall year, it had 48 storms, down from 56 in 1962. But rainfall in 1962 was low. So rainfall is determined by both the number of storms and the amount of rainfall in a single storm.

Section 3

Floyd A. Huff described storm and distribution characteristics, in addition to the above-mentioned factors, geographic location, surrounding terrain and various factors that affect rainfall frequency. He also mentioned nine midwestern states that compose the Midwestern Climate Center, and employing more hydrologic community and other users of storm rainfall information to get more precise mesh information. According to Huff used 11 years data, I think I need more data such as continuously updated data for more than 5 years, or rainfall magnitude and frequency data from other states as the same type of reference to better fit the Gamma model.

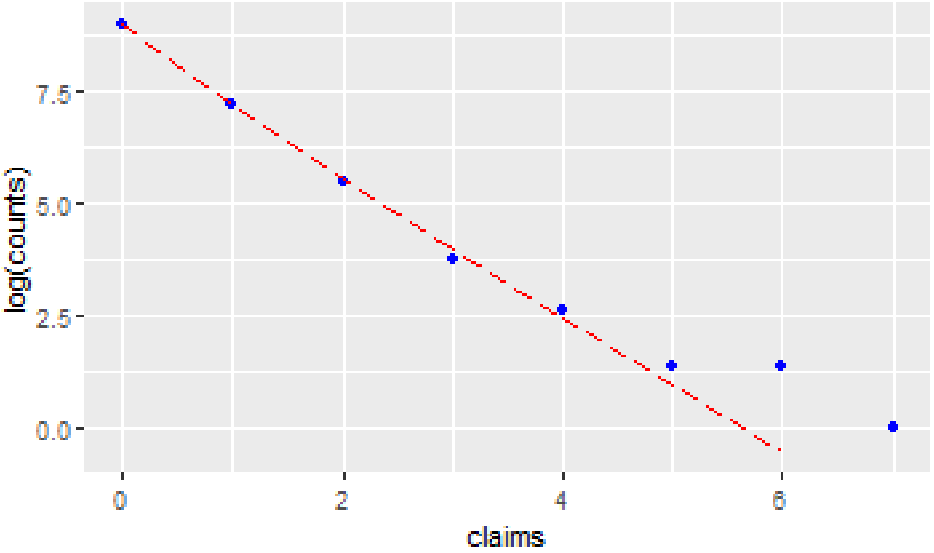
Introduction to Empirical Bayes

Insurance Claims

```
# input data
df <- data.frame(claims=seq(0,7), counts=c(7840,1317,239,42,14,4,4,1))

# Gamma with MLE
f <- function(x,mu,sigma){
  gamma = sigma / (1 + sigma)
  numer = gamma ^ (mu + x) * gamma(mu + x)
  denom = sigma ^ mu * gamma(mu) * factorial(x)
  return(numer/denom)
}

neglog <- function(param){
  mu = param[1]
  sigma = param[2]
  result = -sum(df$counts * log(f(df$claims, mu=mu, sigma=sigma)))
  return(result)
}

# Robbins
Robbins <- 0
for (i in 1:length(df$counts)){
  Robbins[i] <- round(df$claims[i+1] * (df$counts[i+1]/df$counts[i]), 3)
}
df <- cbind(df, Robbins)

# parameters
p <- matrix(c(0.5, 1), 2, 1)
ans_auto <- nlm(f = neglog, p, hessian=T)
mu = ans_auto$estimate[1]
sigma <- ans_auto$estimate[2]
GammaMLE <- (seq(0,6)+1) * f(seq(0,6)+1, mu, sigma)/f(seq(0,6), mu, sigma)
df2 <- t(df)
rbind(df2, GammaMLE)
```

```
##                  [,1]         [,2]        [,3]       [,4]      [,5]     [,6]
## claims      0.0000000    1.0000000   2.0000000  3.0000000  4.000000 5.000000
## counts   7840.0000000 1317.0000000 239.0000000 42.0000000 14.000000 4.000000
## Robbins     0.1680000    0.3630000   0.5270000  1.3330000  1.429000 6.000000
## GammaMLE    0.1641855    0.3982276   0.6322698  0.8663119  1.100354 1.334396
##              [,7]      [,8]
## claims   6.000000 7.0000000
## counts   4.000000 1.0000000
## Robbins  1.750000        NA
## GammaMLE 1.568438 0.1641855
```

```
# plot
df$prediction <- c(f(seq(0,6), mu, sigma) * sum(df$counts), NA)
ggplot(data = df) +
  geom_point(aes(x = claims,y = log(counts)), color='blue') +
  geom_line(aes(x = claims,y = log(prediction)), color='red', lty=5)
```



Missing Species

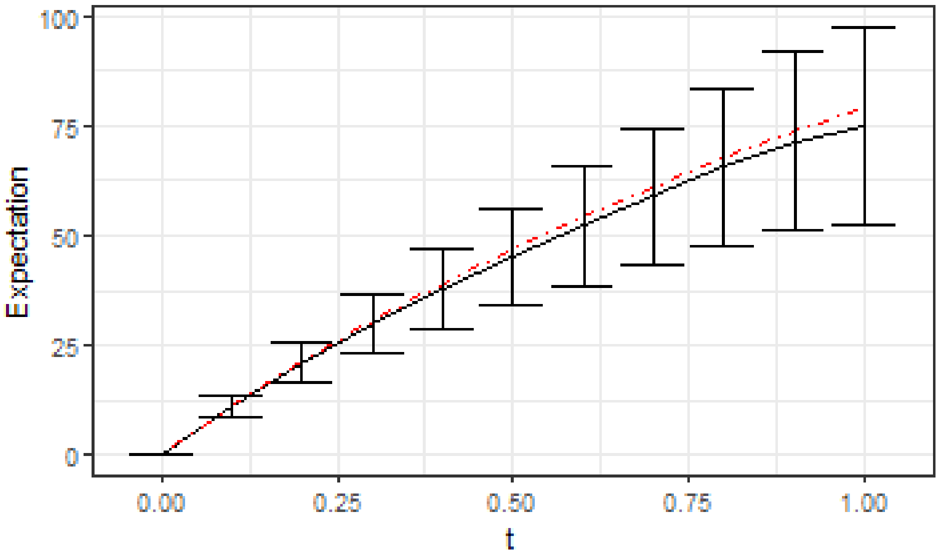
```
x <- seq(1,24)
y <- c(118,74,44,24,29,22,20,19,20,15,12,14,6,12,6,9,9,6,10,10,11,5,3,3)
butterfly <- data.frame(x,y)

# Expectation and Sd
t <- seq(0, 1, by=0.1)
Expectation <- 0
Sd <- 0
for (i in 1:length(t)){
  Expectation[i] <- round(sum(y*(t[i]^x)*(-1)^(x-1)),2)
  Sd[i] <- round(sqrt(sum(y*t[i]^(2))),2)
}
df <- data.frame(t=t, Expectation=Expectation, Sd=Sd)
df
```

```
##      t Expectation    Sd
## 1  0.0        0.00  0.00
## 2  0.1       11.10  2.24
## 3  0.2       20.96  4.48
## 4  0.3       29.79  6.71
## 5  0.4       37.79  8.95
## 6  0.5       45.17 11.19
## 7  0.6       52.15 13.43
## 8  0.7       58.93 15.67
## 9  0.8       65.57 17.91
## 10 0.9       71.56 20.14
## 11 1.0       75.00 22.38
```

```
# parameters
v <- 0.104
sigma <- 89.79
gamma <- sigma / (1 + sigma)
e1 <- 118

# gamma estimate
GammaEstimate <- NULL
for (i in 1:length(t)){
  GammaEstimate[i] <- round(e1*((1 - (1+gamma*t[i])^(-v)) / (gamma * v)),2)
}
GammaEstimate
```

```
##  [1]  0.00 11.20 21.33 30.59 39.09 46.95 54.26 61.08 67.48 73.50 79.18
```

```
# plot
ggplot(data = df) +
  geom_line(aes(x=t,y=Expectation)) +
  geom_line(aes(x=t,y=GammaEstimate), lty=4, color='red') +
  geom_errorbar(aes(x=t, ymin=(Expectation-Sd), ymax=(Expectation+Sd))) +
  theme_bw()
```



What did you learn? What will you do next? What will you do differently as you move forward?

In this project, I learned Gamma MLE, MSE, Order Statistics, profile likelihood, gained a deeper understanding of the application of Empirical Bayes, and judged which model should be used based on distribution and QQ plots.

“In All Likelihood” and “Introduction to Empirical Bayesian” are two very valuable books. Due to limited time, I will continue to read them in the summer vacation, and also try to finish the Shaksapre and Medical

Example in the second book. Of course, I will continue to consolidate and digest the knowledge in Class 677.

When translating the examples in "Introduction to Empirical Bayesian" to R, I ran into a lot of technical problems, like how to code formulas. So just learning the knowledge from the textbook is not enough, it needs more practice.

Reference

1. Order Statistics: https://stackoverflow.com/questions/24211595/order-statistics-in-r?msclkid=fd6683dac56711ecbfcea9bd8a172395
2. Yuli Jin: https://github.com/MA615-Yuli/MA677_final
3. Shuting Li: https://github.com/lst9/MA677-Final-Project
4. Profile Likelihood: https://www.r-bloggers.com/2015/11/profile-likelihood/
5. Gamma QQ Plot: https://stackoverflow.com/questions/50092506/how-to-draw-a-qq-plot-in-r