rainGauge_midterm.R

lenovo

Sun Oct 18 22:04:55 2015

```r
#input raw rain gauge data
library(stringr)
theFiles <- dir("raw/",pattern="\\.txt")
theFiles
```

```
##  [1] "L-00-01.txt" "L-00-02.txt" "L-00-03.txt" "L-00-04.txt" "L-00-05.txt"
##  [6] "L-00-06.txt" "L-00-07.txt" "L-00-08.txt" "L-00-09.txt" "L-00-10.txt"
## [11] "L-00-11.txt" "L-00-12.txt" "L-01-01.txt" "L-01-02.txt" "L-01-03.txt"
## [16] "L-01-04.txt" "L-01-05.txt" "L-01-06.txt" "L-01-07.txt" "L-01-08.txt"
## [21] "L-01-09.txt" "L-01-10.txt" "L-01-11.txt" "L-01-12.txt" "L-02-01.txt"
## [26] "L-02-02.txt" "L-02-03.txt" "L-02-04.txt" "L-02-05.txt" "L-02-06.txt"
## [31] "L-02-07.txt" "L-02-08.txt" "L-02-09.txt" "L-02-10.txt" "L-02-11.txt"
## [36] "L-02-12.txt" "L-03-01.txt" "L-03-02.txt" "L-03-03.txt" "L-03-04.txt"
## [41] "L-03-05.txt" "L-03-06.txt" "L-03-07.txt" "L-03-08.txt" "L-03-09.txt"
## [46] "L-03-10.txt" "L-03-11.txt" "L-03-12.txt" "L-04-01.txt" "L-04-02.txt"
## [51] "L-04-03.txt" "L-04-04.txt" "L-04-05.txt" "L-04-06.txt" "L-04-07.txt"
## [56] "L-04-08.txt" "L-04-09.txt" "L-04-10.txt" "L-04-11.txt" "L-04-12.txt"
```

```r
for (a in theFiles){
  nameToUse <- str_sub(string=a,start=1,end=7)
  temp <- read.csv(file=file.path("raw",a), skip = 2, stringsAsFactors = F)
  assign(x=nameToUse,value=temp)
}
L1 <- rbind(`L-00-01`, `L-00-02`, `L-00-03`, `L-00-04`, `L-00-05`, `L-00-06`, `L-00-07`, `L-00-08`,
            `L-00-09`, `L-00-10`, `L-00-11`, `L-00-12`, `L-01-01`, `L-01-02`, `L-01-03`, `L-01-04`,
            `L-01-05`, `L-01-06`, `L-01-07`, `L-01-08`, `L-01-09`, `L-01-10`, `L-01-11`, `L-01-12`,
            `L-02-01`, `L-02-02`, `L-02-03`, `L-02-04`, `L-02-05`, `L-02-06`, `L-02-07`, `L-02-08`,
            `L-02-09`, `L-02-10`, `L-02-11`, `L-02-12`, `L-03-01`, `L-03-02`, `L-03-03`, `L-03-04`,
            `L-03-05`, `L-03-06`, `L-03-07`, `L-03-08`, `L-03-09`, `L-03-10`, `L-03-11`, `L-03-12`,
            `L-04-01`, `L-04-02`, `L-04-03`, `L-04-04`, `L-04-05`, `L-04-06`, `L-04-07`, `L-04-08`,
            `L-04-09`, `L-04-10`, `L-04-11`, `L-04-12`)

#data cleaning
L1 <- L1[,2:length(L1)]
colnames(L1) <- c(1:24)
UL <- as.vector(t(L1))
# "T" represents trace; "----" represents no rain;
#I don't know what "M" means, so I regard it as no rain;
#In order to get rain storm data easily, "T" is translated as 0
#and no rain, both "----" and "M", are translated as -1
UL[UL == "T   "] <- "0"
UL[UL == "----"] <- "-1"
UL[UL == "M   "] <- "-1"

rainGauge <- as.numeric(UL)
head(rainGauge)
```

```
## [1] -1 -1 -1 -1 -1 -1
```

```r
#create function storm to get rain storm data
storm <- function(x) {
# x should be a vector
  i = 1
  while(i < length(x)+1){
    tmp = 0
    while((x[i] >= 0)){
    tmp <- tmp + x[i]
    i = i+1
    }
    sum <- c(sum, tmp)
    if((x[i] < 0))
    {
    i = i + 1
    }
  }
    return(sum)
  }


rain <- storm(rainGauge)
head(rain)
```

```
## [[1]]
## function (..., na.rm = FALSE)  .Primitive("sum")
##
## [[2]]
## [1] 0
##
## [[3]]
## [1] 0
##
## [[4]]
## [1] 0
##
## [[5]]
## [1] 0
##
## [[6]]
## [1] 0
```

```r
rain1 <- as.data.frame(unlist(rain[c(2:length(rain))]))
rain2 <- rain1[rain1 != 0]

#rain2 is rain storm data
head(rain2)
```

```
## [1] 0.03 0.03 0.01 0.01 0.97 0.06
```

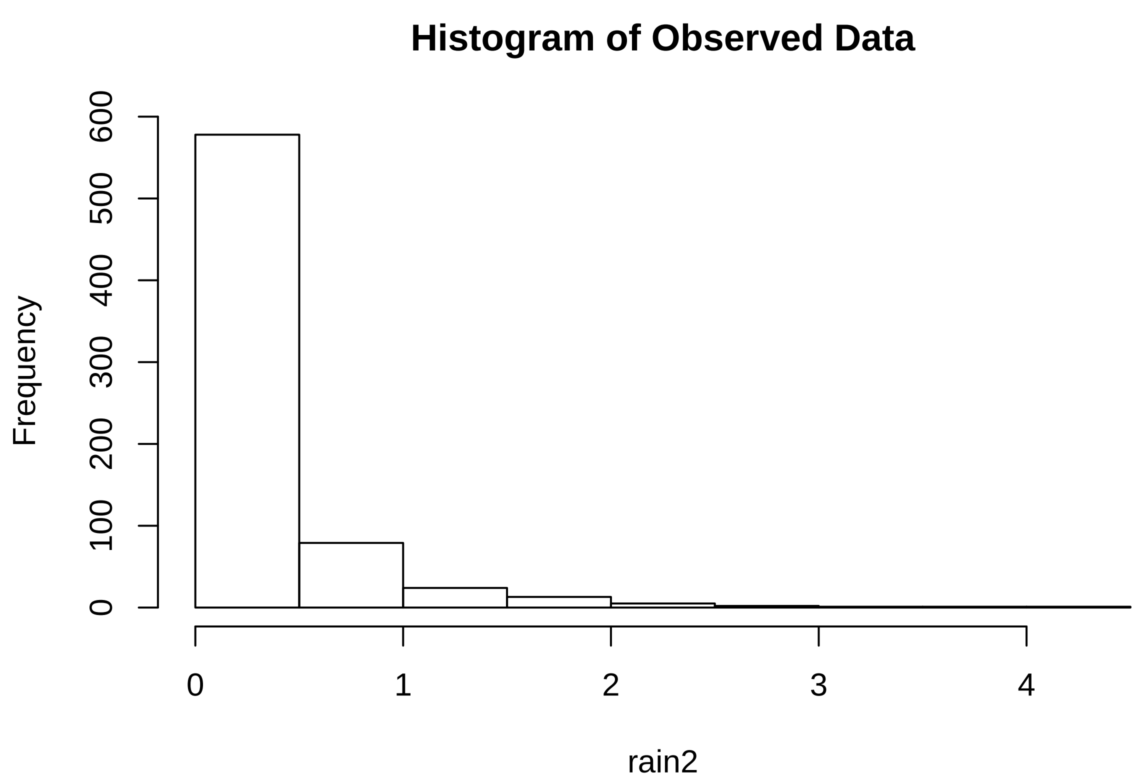
```r
#fitting distribution

#EDA----mean and variance
storm.mean <- mean(rain2)
storm.mean
```

```
## [1] 0.2831108
```

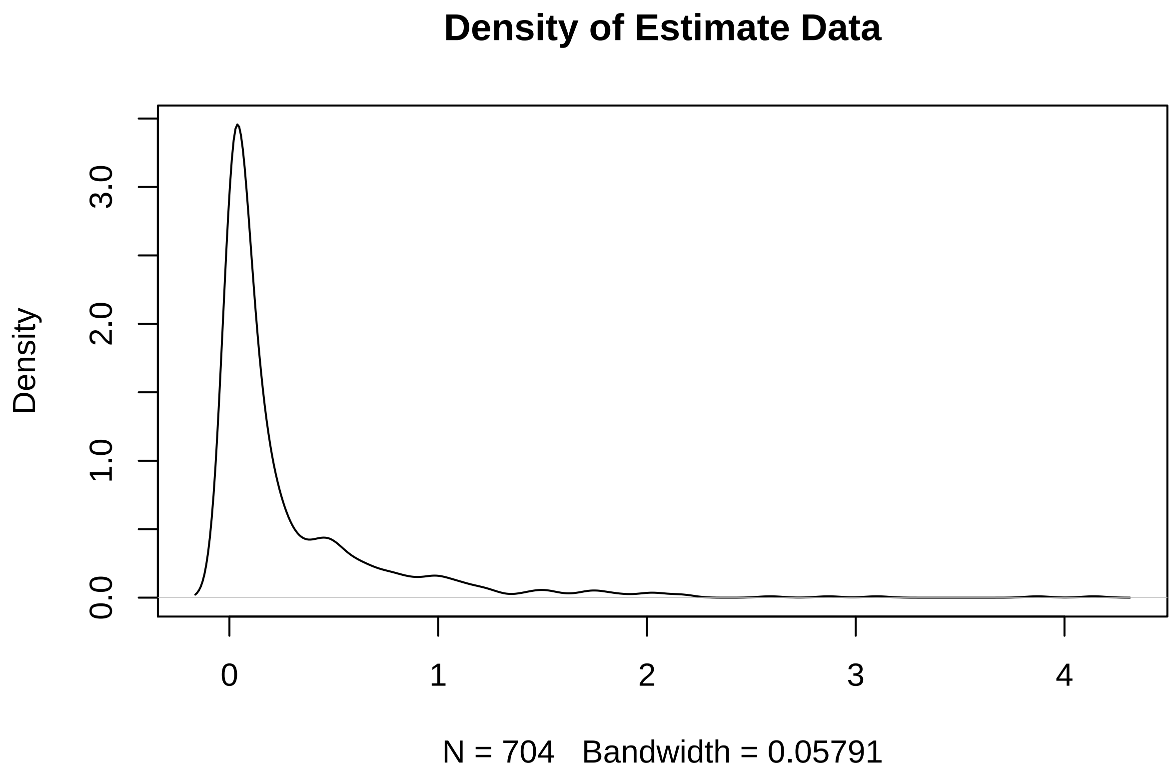
```r
storm.var <- var(rain2)
storm.var
```

```
## [1] 0.2218382
```

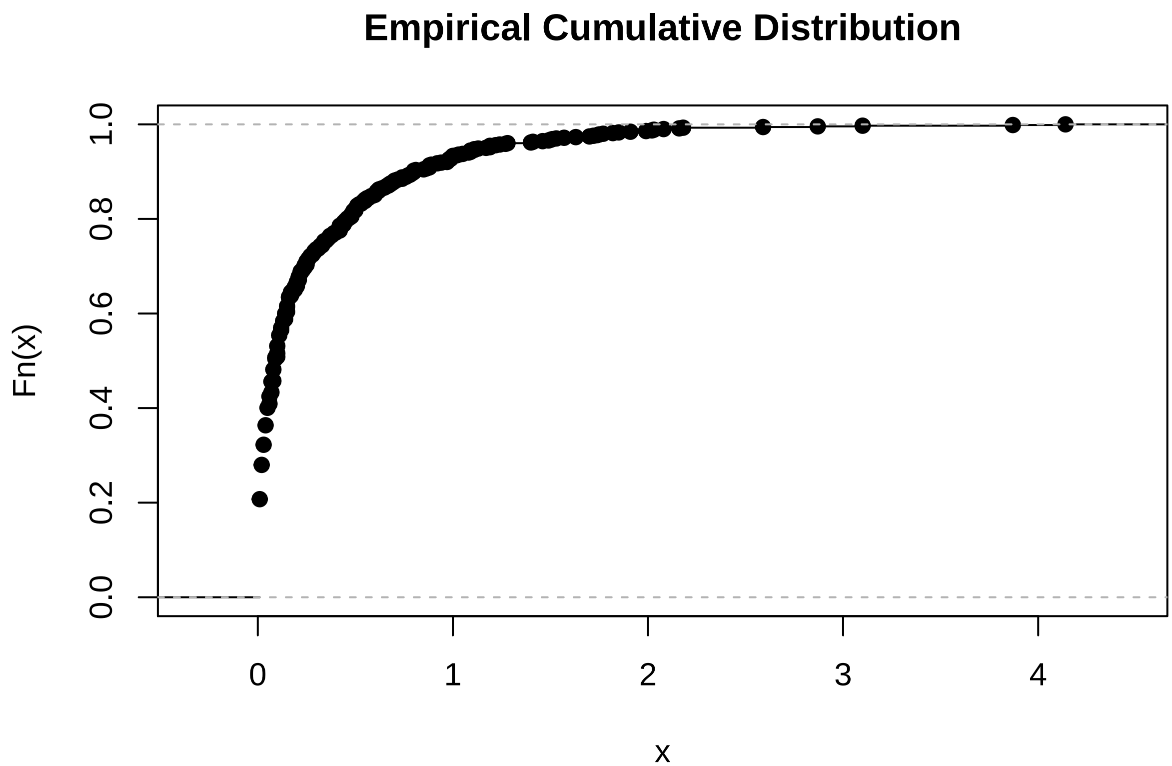
```r
#graphs
library(ggplot2)
#get histogram
hist(rain2, main = "Histogram of Observed Data")
```




```r
#estimate frequency density
plot(density(rain2), main = "Density of Estimate Data")
```




```r
#get ecdf plot
plot(ecdf(rain2), main = "Empirical Cumulative Distribution")
```




```
#assume rain storm data have gamma distribution,
#use MLE to get the estimation of parameters
#then use qqplot() tests whether this hyphothesis is right or not
#lambda and alpha are estimated by the method of moments, which will be used as the start point of MLE
lambda <- storm.mean/storm.var
lambda
```

```
## [1] 1.276204
```

```
alpha <- (storm.mean^2)/storm.var
alpha
```

```
## [1] 0.3613071
```

```
#maximum likelihood estimates
n <- length(rain2)

minus.likelihood <- function(theta) {-(n*theta[1]*log(theta[2])-n*lgamma(theta[1])+(theta[1]-1)*sum(log

max.likelihood <- nlminb(start=c(alpha, lambda), obj = minus.likelihood)

max.likelihood$par
```

```
## [1] 0.5461541 1.9291179
```

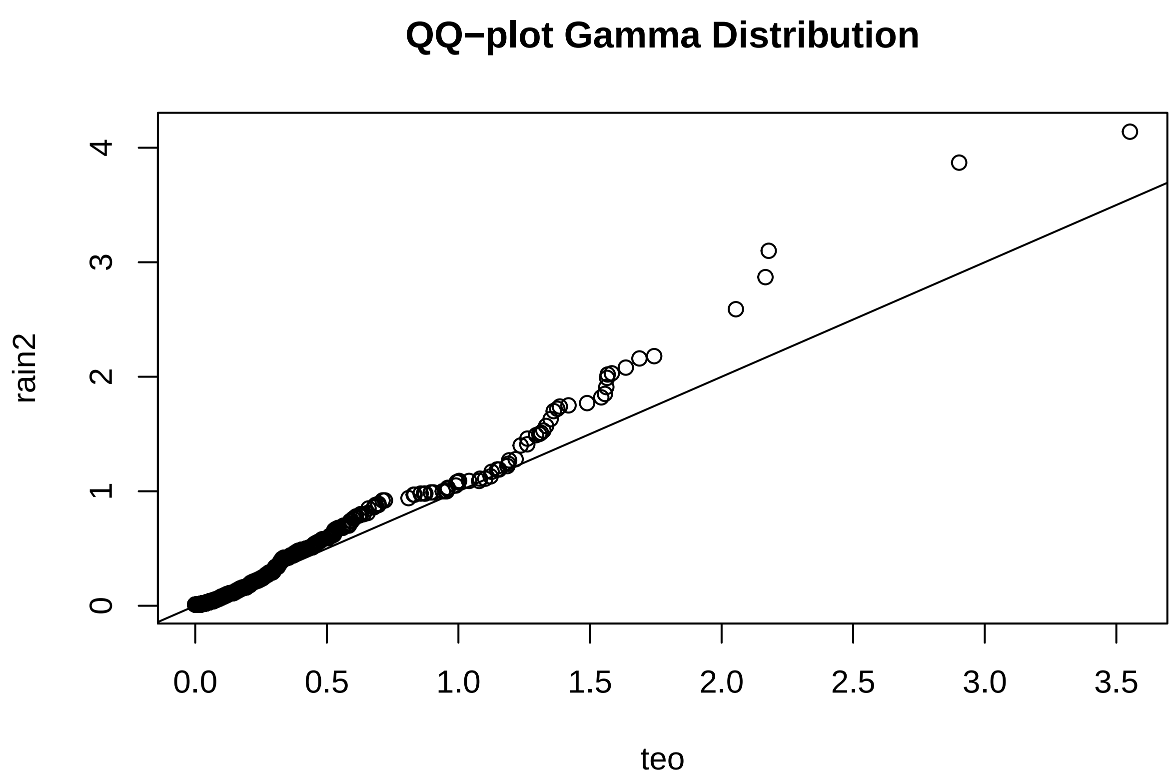
```
alpha1 <- max.likelihood$par[1]
alpha1
```

```
## [1] 0.5461541
```

```
lambda1 <- max.likelihood$par[2]
lambda1
```

```
## [1] 1.929118
```

```
#QQ-plot
set.seed(50)
teo <- rgamma(length(rain2), shape = alpha1, rate = lambda1)#theorotical gamma distribution
qqplot(teo, rain2, main = "QQ-plot Gamma Distribution")
abline(0,1)# reference line
```




```
#about 1/2 rain storm data point fall approximately along the reference line
#So probably logan airport rain storm data don't follow gamma distribution

#ks.test
ks.test(rain2, teo)
```

```
## Warning in ks.test(rain2, teo): p-value will be approximate in the presence
## of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  rain2 and teo
## D = 0.14205, p-value = 1.355e-06
## alternative hypothesis: two-sided
```

```
#Because p-value is below 0.001, we reject the null hypothesis.
#So the rain storm data do not follow gamma distribution
```