677_fp

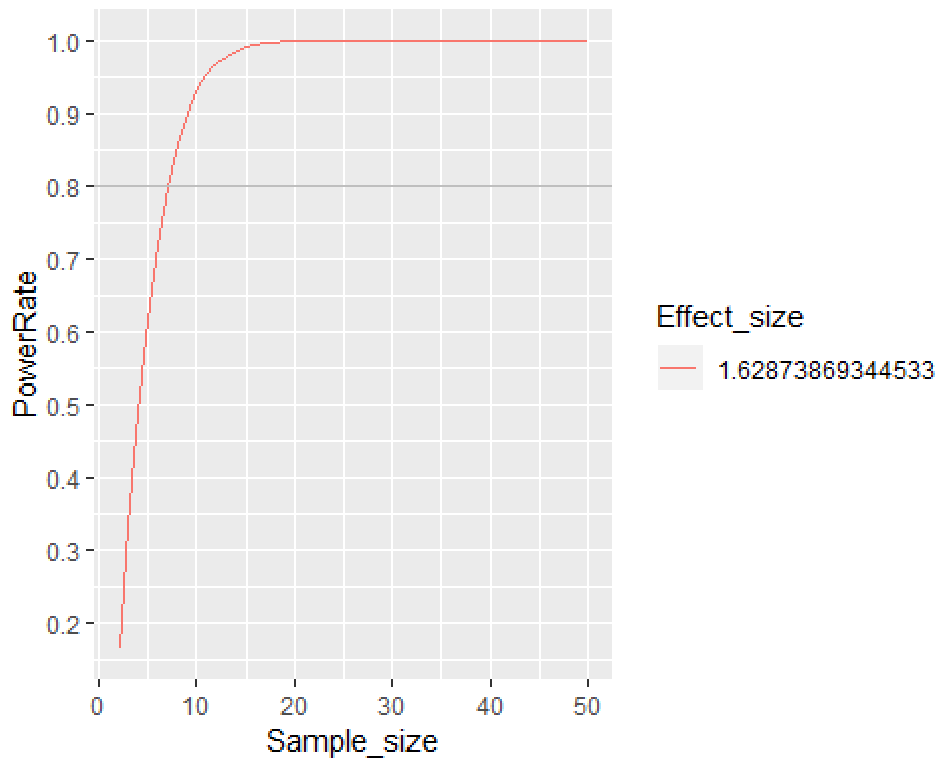
Samuel

May 7, 2019

1 Statistics and the Law

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.
2.1 --

## v ggplot2 3.0.0       v purrr   0.2.5
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.1       v stringr 1.3.1
## v readr   1.1.1       v forcats 0.3.0

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(readr)
library(dplyr)
library(pwr)
library(fitdistrplus)

## Warning: package 'fitdistrplus' was built under R version 3.5.3

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 3.5.2

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: survival

## Loading required package: npsurv

## Warning: package 'npsurv' was built under R version 3.5.2
```

```
## Loading required package: lsei

## Warning: package 'lsei' was built under R version 3.5.2

library(MASS)


acorn <- read_csv("acorn.csv")

## Parsed with column specification:
## cols(
##   BANK = col_character(),
##   MIN = col_double(),
##   WHITE = col_double(),
##   HIMIN = col_double(),
##   HIWHITE = col_double()
## )

test1 <- t.test(acorn$MIN, acorn$WHITE, paired = T)
test1$p.value

## [1] 5.619188e-10

#According to the p-value, the refusal rate for minority applicants is signif
icantly different from the refusal rate for white applicants which warrants a
corrective action.

#Further Analysis, Sufficient for this dataset.
#power analysis
##Settings
pa <- cbind(NULL)
n <- seq(2, 50, by = 1)
##Loop
for (i in n) {
  papwr <- pwr.t2n.test(
    n1 = i, n2 = i,
    sig.level = 0.05, power = NULL,
    d = abs(mean(acorn$MIN)-mean(acorn$WHITE))/sd(acorn$MIN), alternative = "
two.sided"
  )
  pa <- rbind(pa, papwr$power)
}

pa <- as.data.frame(pa)

ggplot(pa) + geom_line(aes(x = n, y = V1, color = "salmon")) + scale_color_di
screte(name = "Effect_size", labels = c(abs(mean(acorn$MIN)-mean(acorn$WHITE)
)/sd(acorn$MIN))) + geom_hline(yintercept = 0.8, color = "gray") + scale_y_co
ntinuous(breaks = seq(0, 1, by = 0.1)) + xlab("Sample_size") + ylab("PowerRat
e")
```



According to the power plot, the sample size that needed to reach 80% power rate is around 7, but we have total of 20 sample size in the acorn dataset, which means that it is sufficient.

2 Comparing supplies

```
#H0:All schools have the same quality
#H1:All schools have the different quality

product <- matrix(c(12,23,89,8,12,62,21,30,119), byrow=TRUE, ncol=3, nrow = 3
)
colnames(product) <- c("dead","art","fly")
product <- as.data.frame(product)
product$schname[1] <- "Area51"
product$schname[2] <- "BDV"
product$schname[3] <- "Giffen"
chisq.test(product[,1:3],correct = F)

##
##  Pearson's Chi-squared test
##
## data:  product[, 1:3]
## X-squared = 1.3006, df = 4, p-value = 0.8613
```

According to the Chi-squared test, this p-value is 0.8613 which is not significant. Thus, we fail to reject the null hypothesis that all schools have the same quality.

3 How deadly are sharks?

```
shark <- read_csv("sharkattack.csv")

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   Date = col_character(),
##   Country = col_character(),
##   `Country code` = col_character(),
##   Type = col_character(),
##   Continent = col_character(),
##   Hemisphere = col_character(),
##   Activity = col_character(),
##   Fatal = col_character()
## )

shark_US <- shark[which(shark$`Country code` == "US"),]
shark_AU <- shark[which(shark$`Country code`== "AU"),]
shark_new <- rbind(shark_US, shark_AU)
shark_new$X1 <- NULL

fatal_count <- shark_new %>% group_by(shark_new$`Country code`, shark_new$Fat
al) %>% summarize(count = n())

fatal_count <- fatal_count[-which(fatal_count$`shark_new$Fatal` == "UNKNOWN")
, ]
fatal_count

## # A tibble: 4 x 3
## # Groups:   shark_new$`Country code` [2]
##   `shark_new$\`Country code\`` `shark_new$Fatal` count
##   <chr>                        <chr>             <int>
## 1 AU                           N                   879
## 2 AU                           Y                   318
## 3 US                           N                  1795
## 4 US                           Y                   217

fatal_count_N <- fatal_count[which(fatal_count$`shark_new$Fatal` == "N"),]
fatal_count_Y <- fatal_count[which(fatal_count$`shark_new$Fatal` == "Y"),]
fatal_count <- left_join(fatal_count_N, fatal_count_Y, by = "shark_new$`Count
ry code`")
fatal_count$total <- fatal_count$count.x + fatal_count$count.y
fatal_count$fatal_percentage <- fatal_count$count.y/ fatal_count$total
colnames(fatal_count) <- c("countrycode", "Non_Fatal", "Non_Fatal_count", "Fa
tal", "Fatal_count", "total_count","Fatal_percentage")
fatal_count$Non_Fatal <- NULL
fatal_count$Fatal <- NULL
fatal_count
```

```
## # A tibble: 2 x 5
## # Groups:   shark_new$`Country code` [2]
##   countrycode Non_Fatal_count Fatal_count total_count Fatal_percentage
##   <chr>                 <int>       <int>       <int>            <dbl>
## 1 AU                      879         318        1197            0.266
## 2 US                     1795         217        2012            0.108
```

From this table, we could see that although the number of shark attack reports in US is more than that in Austrilia, the percentage of fatal attack in Austrilia is much more than percentage of fatal attack in US.

4 power analysis

Since the books says the power that detects for the difference between hypothetical parameters .65 and .45 is .48 and the power that detects for the difference between hypothetical parameters .25 and .05 is .82. However, the difference between both pairs of values is .20. It means that hypothetical parameters of this binomial distribution doesn not provide a scale of equal units of detectability.On the other hand, when using arcsine transformation, it transforms the scale of proportional parameter to the scale from $-\pi/2$ to $\pi/2$. In addition, t1 -t2 = h, which provode a scale of euqal dectectability. Thus, it is a solution to solve the problem that does not provide of equal units of detectability.

5 Estimators

Exponential

Exponential

$$f(x,\lambda) = \lambda e^{-\lambda x} \Rightarrow E[X] = \int_0^\infty x \lambda e^{-\lambda x}\, dx$$

$$= -\frac{x e^{-\lambda x}}{\lambda}\Big|_0^\infty - -\frac{e^{-\lambda x}}{\lambda x}\Big|_0^\infty = \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty = \frac{1}{\lambda}$$

$$E[X] = \bar{x} \Rightarrow \frac{1}{\lambda} \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}.$$

$$\therefore L(\lambda, x_1, \dots, x_n) = \lambda^n e^{-\lambda \sum_{i=0}^n x_i}$$

$$\Rightarrow l(\lambda, x_1, \dots, x_n) = n\log(\lambda) - \lambda \sum_{i=0}^n x_i \Rightarrow \frac{\Delta l}{\Delta \lambda} = \frac{n}{\lambda} - \sum_{i=0}^n x_i = 0$$

$$\Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=0}^n x_i} = \frac{1}{\bar{x}}.$$

A new distribution

#2 Distribution.

$$E(X) = \int_0^1 x(1-\theta+2\theta x)\,dx = \frac{1}{2} - \frac{1}{2}\theta + \frac{2}{3}\theta = \frac{1}{2} + \frac{\theta}{6}$$

$$\therefore \quad \bar{x} = \frac{1}{2} + \frac{1}{6}\theta \Rightarrow \hat{\theta} = 6\bar{x} - 3$$

$$L(\theta, x_1 \cdots x_n) = \prod_{i=1}^{n} (1-\theta+2\theta x_i)$$

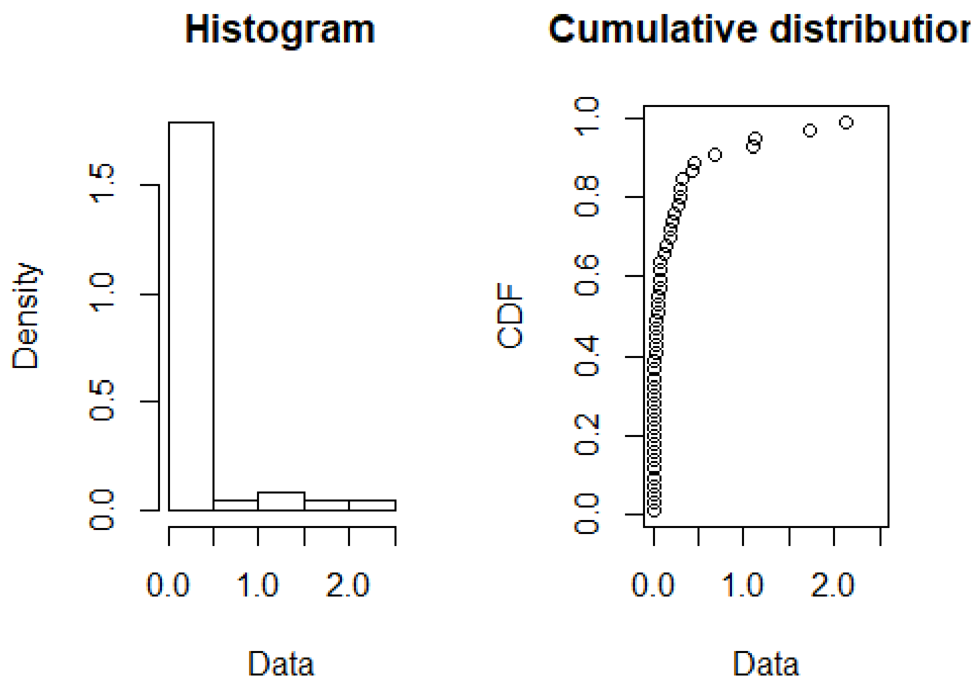
$$l(\theta, x_1, \cdots x_n) = \sum_{i=1}^{n} \ln(1-\theta+2\theta x_i)$$

$$\Rightarrow \frac{\partial l}{\partial \theta} = \sum_{i=1}^{n} \frac{2x_i - 1}{1-\theta+2\theta x_i} = 0.$$

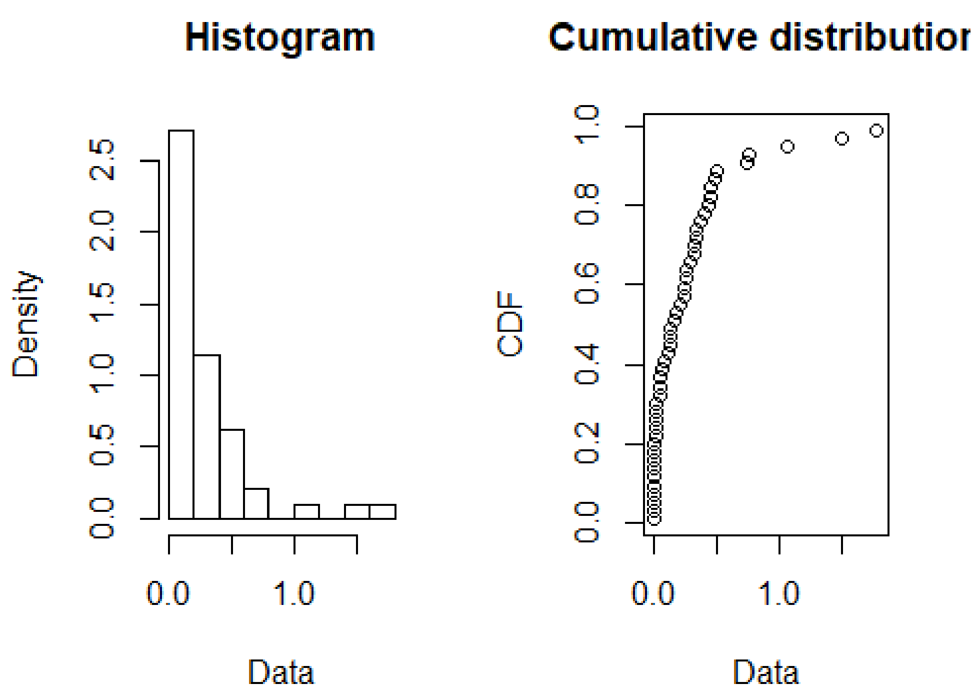
Thus, we could derive the $\hat{\theta}$ result.

Rain in Southern Illinois

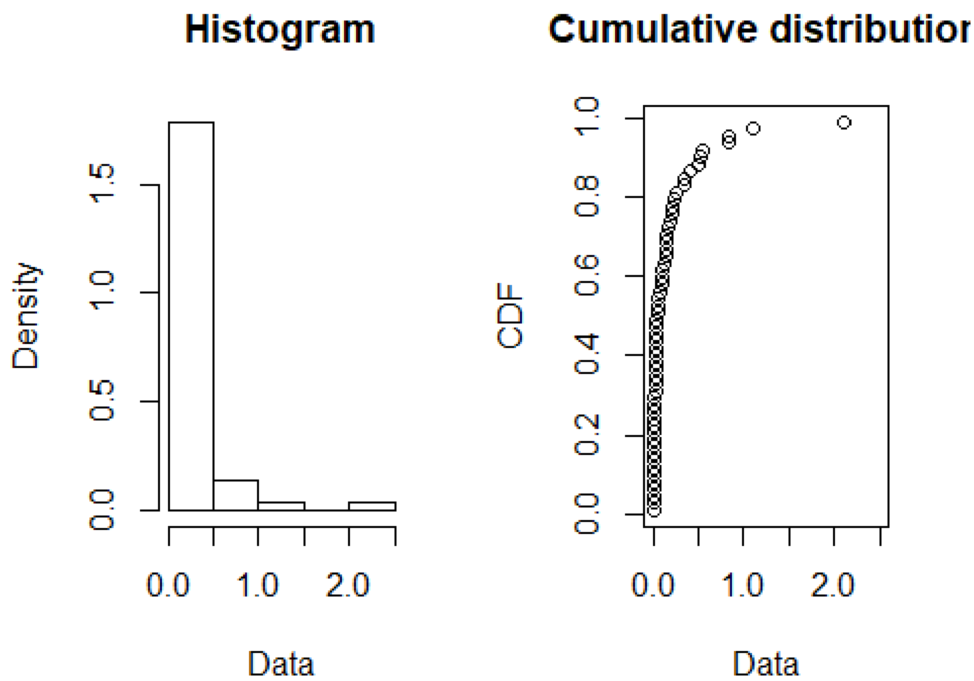
```
rain60 <- read.table("ill-60.txt")
rain61 <- read.table("ill-61.txt")
rain62 <- read.table("ill-62.txt")
rain63 <- read.table("ill-63.txt")
rain64 <- read.table("ill-64.txt")

rain60 <- as.numeric(as.array(rain60$V1))
rain61 <- as.numeric(as.array(rain61$V1))
rain62 <- as.numeric(as.array(rain62$V1))
rain63 <- as.numeric(as.array(rain63$V1))
rain64 <- as.numeric(as.array(rain64$V1))

plotdist(rain60)
```



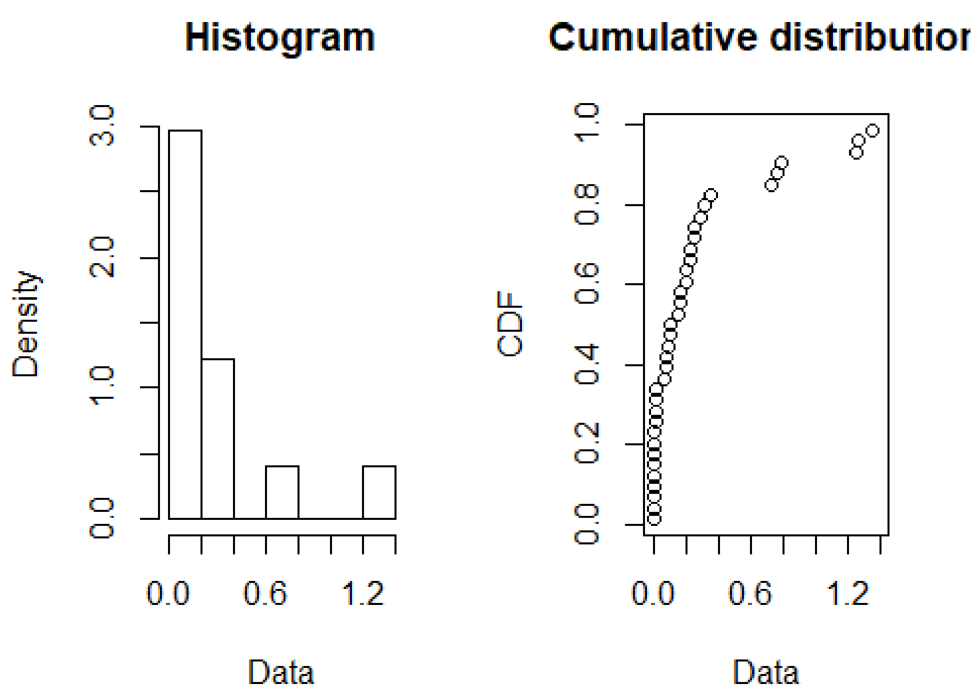

```
plotdist(rain61)
```



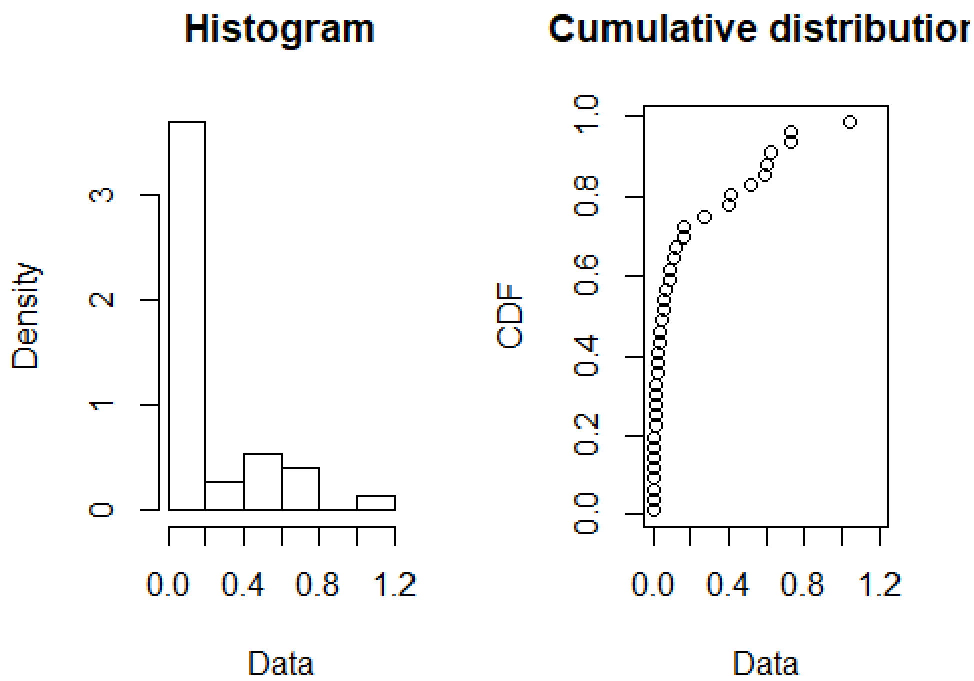

```
plotdist(rain62)
```



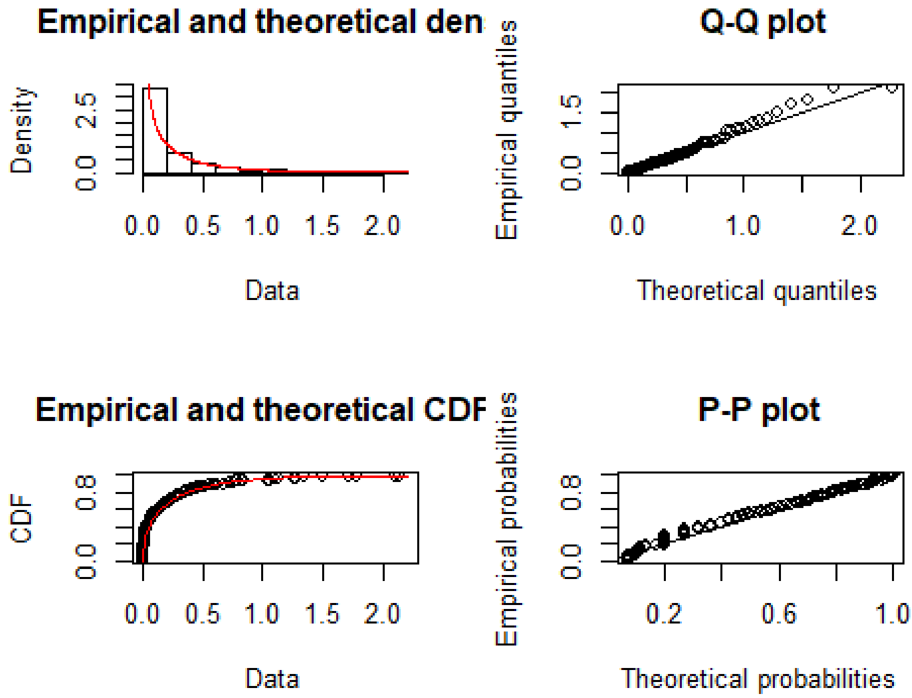

```
plotdist(rain63)
```




```
plotdist(rain64)
```




```
totalrain <- as.data.frame(t(c(sum(rain60),sum(rain61),sum(rain62),sum(rain63
),sum(rain64))))
colnames(totalrain)<-c("Total1960","Total1961","Total1962","Total1963","Total
1964")
#1961 has the most rainfall, and then the rainfall decreases each year. Until
1964, the rainfall become the least.

#Gamma distribution
rain01234 <- c(rain60,rain61,rain62,rain63,rain64)
raingamma <- fitdist(rain01234, "gamma")
plot(raingamma)
```




```
#Based on the plot, I am agree that gamma distribution fits well. From Q-Q pl
ot, we could see most of the points are distributed near the line.

rainmom <- fitdist(rain01234, "gamma",method = "mme")
rainmle <- fitdist(rain01234, "gamma",method = "mle")
bootmom <- bootdist(rainmom)
bootmle <- bootdist(rainmle)
summary(bootmom)

## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.3909681 0.2757808 0.5269659
## rate  1.7429150 1.1551499 2.6426050

summary(bootmle)

## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4423184 0.3841021 0.5166918
## rate  1.9796765 1.5601977 2.5147024

#We can see for bootstrap, the 95% CI interval is (0.28,0.52) as well as rate
is (1.19,2.56). On the other hand, for MLE, the 95% CI interval is (0.38,0.52
) as well as rate is (1.56,2.55).In this case, I would prefer MLE as my estim
ator since it has a smaller CI interval result in a lower variance
```

6 Analysis of decision theory article

#6 for $\delta \in [0,1]$.

$$V(\delta, P) = E[y(A)](1-\delta) + E[y(B)]\delta$$
$$= \alpha + (-\alpha + \beta)\delta. \quad (\alpha = E[y(A)],\ \beta = E[y(B)])$$
$$\Rightarrow W(\delta, P, Q) = \alpha + (\beta - \alpha) E[\delta(\psi)]. \quad (V(P, P, \psi) = \alpha + (\beta - \alpha + P)\delta(\psi))$$

if probability of success $\alpha = P[y(A)=1]$, not success $\beta = P[y(B)=1]$.

$$\therefore \cancel{E[\delta}\ W(\delta, P, N) = \alpha + (-\alpha + \beta) E[\delta(n)].$$
$$\therefore E[\delta(n)] = \sum_{i=0}^{N} \delta(i) \underbrace{f(\cancel{A}\, n, \beta, N)}_{\text{probability of success.}}$$

$$\begin{cases} \delta(n) = 0, & n < n_0 \\ \delta(n) = 1, & n > n_0 \\ \delta(n) = \lambda, & n = n_0 \end{cases} \quad (\text{for } 0 \le n_0 \le N,\ 0 \le \lambda \le 1)$$

If let $(P_s, s \in S) = (0,1)$ with parameter (c,d).

By Bayes rules, $\delta(n) = 0$ for $(c+n)/(c+d+N) < \alpha$

$\delta(n) = 1$ for $(c+n)/(c+d+N) > \alpha$

$\delta(n) = \lambda$ for $(c+n)/(c+d+N) = \alpha$.