MA677 FINAL PROJECT

Kecheng Liang

May 2019

Statistics and the Law

two sample t-test Null hypothesis: refuse rate of minority applicant is the same as that of white applicant
Alternative hypothesis: refuse rate of minority applicant is higher that of white applicant

```
acorn<-read.csv("acorn.csv")
test1 <- var.test(acorn$MIN,acorn$WHITE)
test1
```

```
##
##  F test to compare two variances
##
## data:  acorn$MIN and acorn$WHITE
## F = 2.8026, num df = 19, denom df = 19, p-value = 0.02993
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.109297 7.080589
## sample estimates:
## ratio of variances
##           2.802583
```

```
test2 <- t.test(acorn$MIN, acorn$WHITE, alternative = "greater",var.equal = FALSE)
test2
```

```
##
##  Welch Two Sample t-test
##
## data:  acorn$MIN and acorn$WHITE
## t = 6.2533, df = 31.028, p-value = 2.979e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  15.49313      Inf
## sample estimates:
## mean of x mean of y
##   36.8815   15.6250
```

First, I did the F test to compare two variances. Since p-value = 0.02993, reject the null hypothesis, the two variances are not same. Then I did the two sample t-test. The result shows that the p-value = 2.979e-07, reject the null hypothesis. The refuse rate of minority applicant is higher that of white applicant.

Comparing Suppliers

Chi-square test Null hypothesis: all three schools produces the same quality
Alternative hypothesis: at least one of these three schools produces different quality

```
df <- matrix(c(12,8,21,23,12,30,89,62,119),nrow = 3,byrow = FALSE)
colnames(df) <- c("Dead Bird","Display Art","Flying Art")
```

```r
rownames(df) <- c("Area 51","BDV","Giffen")
chisq.test(df)
```

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

By doing the chi-square test, p-value = 0.8613 which is greater than 0.05. Therefore, we fail to reject the null hypothesis that all three schools produces the same quality.

How deadly are sharks?

```r
shark<-read.csv("sharkattack.csv")
sharkdf <- shark %>%
            filter(Country == "United States" | Country == "Australia") %>%
            filter(Type == "Provoked" | Type == "Unprovoked")
table <- table(droplevels(sharkdf)$Country,droplevels(sharkdf)$Type)
kable(prop.table(table,margin = 1))
```

	Provoked	Unprovoked
Australia	0.1304791	0.8695209
United States	0.1073826	0.8926174

```r
chisq.test(table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 3.084, df = 1, p-value = 0.07907
```

```r
sharkdf2 <- shark %>%
            filter(Country == "United States" | Country == "Australia") %>%
            filter(Fatal == "Y" | Fatal == "N")
table2 <- table(droplevels(sharkdf2)$Country,droplevels(sharkdf2)$Fatal)
kable(prop.table(table2,margin = 1))
```

	N	Y
Australia	0.7343358	0.2656642
United States	0.8921471	0.1078529

```r
chisq.test(table2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table2
## X-squared = 133.41, df = 1, p-value < 2.2e-16
```

```r
pwr.chisq.test(w = ES.w2(prop.table(table2)), N=879+318+1795+217, df = 1, sig.level = 0.05)
```

```
##
##      Chi squared power calculation
##
##               w = 0.2047583
##               N = 3209
##              df = 1
##       sig.level = 0.05
##           power = 1
##
## NOTE: N is the number of observations
```

I did 2 chi-square tests. The first one is try to compare the provoked and unprovoked between US and Australia. The result shows that p-value = 0.07907, and we do not reject the null hypothesis. Therefore, there is no association betweeen two variables. The p-value from second one is < 2.2e-16. Therefore, we reject the null hypothesis, there is association between fatal and country. The attack in Australia is much more deadly. The power is 1.

Power analysis

In the book, it said that the hypothetical parameters of this binomial distribution doesn not provide a scale of equal units of detectability. Arcsin transformation could solve the problem that falling into one side of the range.symbol = 2 arcsin root(P).

Estimators

Exponential

$$f(x_i;\lambda) = \lambda e^{-\lambda x}$$

mm:

$$E(x) = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \lambda\int_0^\infty x e^{-\lambda x}\,dx = \frac{1}{\lambda}$$

$$\Rightarrow \bar{X} = \frac{1}{\lambda} \quad \Rightarrow \hat{\lambda} = \frac{1}{\bar{X}}$$

MLE =

$$L(\lambda; x_1,\dots,x_n) = f(x_1)f(x_2)f(x_3)\cdots f(x_n)$$

$$= \lambda e^{-\lambda \Sigma x_i}$$

log: $\ell(\lambda; x_1 \dots x_n) = n\log(\lambda) - \lambda \Sigma x_i$

derivative: $\dfrac{d\ell}{d\lambda} = \dfrac{n}{\lambda} - \Sigma x_i = 0$

$$\Rightarrow \frac{n}{\lambda} = \Sigma x_i$$

$$\Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}$$

A new distribution

$$f(x) = \begin{cases} (1-\theta)+2\theta x & 0<x<1 \\ 0 & \text{otherwise.} \end{cases}$$

MM:

$$E(x) = \int_0^1 x((1-\theta)+2\theta x)\,dx$$

$$= (1-\theta)\int_0^1 x\,dx + \int_0^1 2\theta x^2\,dx$$

$$= \frac{1}{2} - \frac{1}{2}\theta + \frac{2}{3}\theta = \frac{1}{2} + \frac{1}{6}\theta$$

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

MLE:

$$L(\theta; x_1 \cdots x_n) = \Pi[(1-\theta)+2\theta x_i]$$

log: $$\mathcal{L}(\theta; x_1, \cdots x_n) = \Sigma \ln[(1-\theta)+2\theta x_i]$$

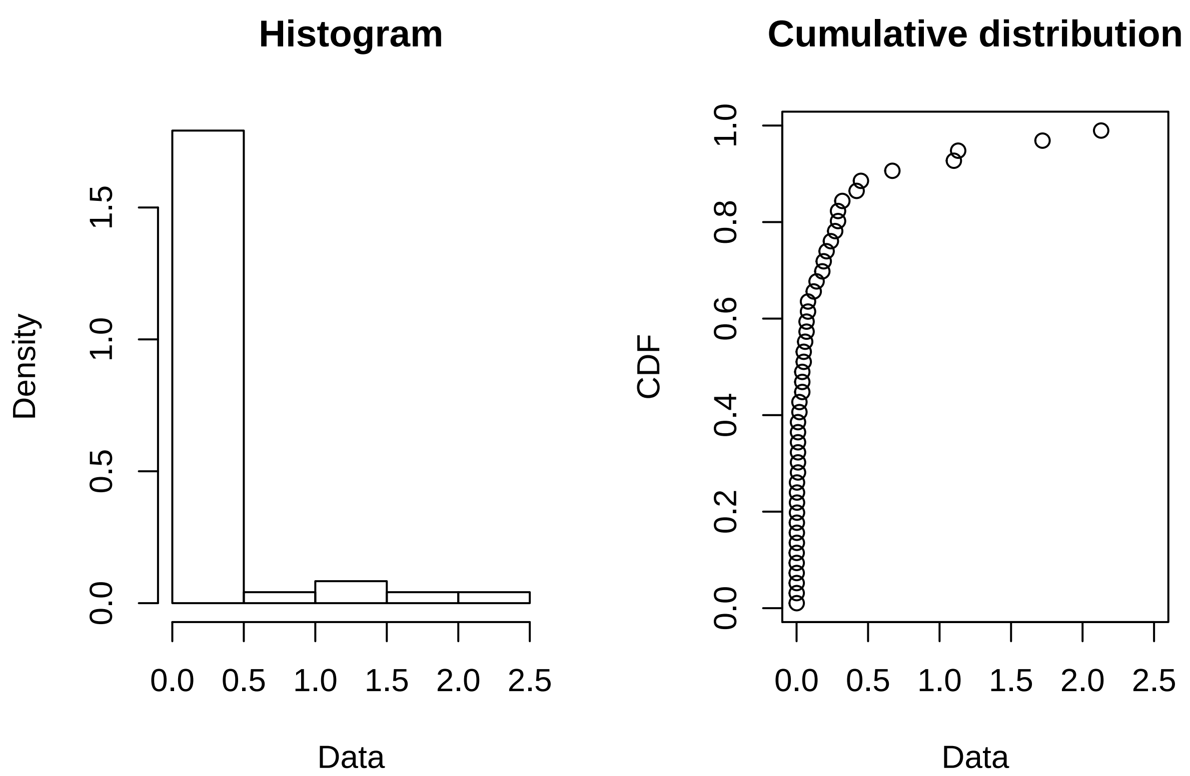
derivative: $$\frac{d\ell}{d\theta} = \Sigma \frac{2x_i - 1}{1-\theta+2\theta x_i} = 0$$

$\hat{\theta}$ = the solution of ↑

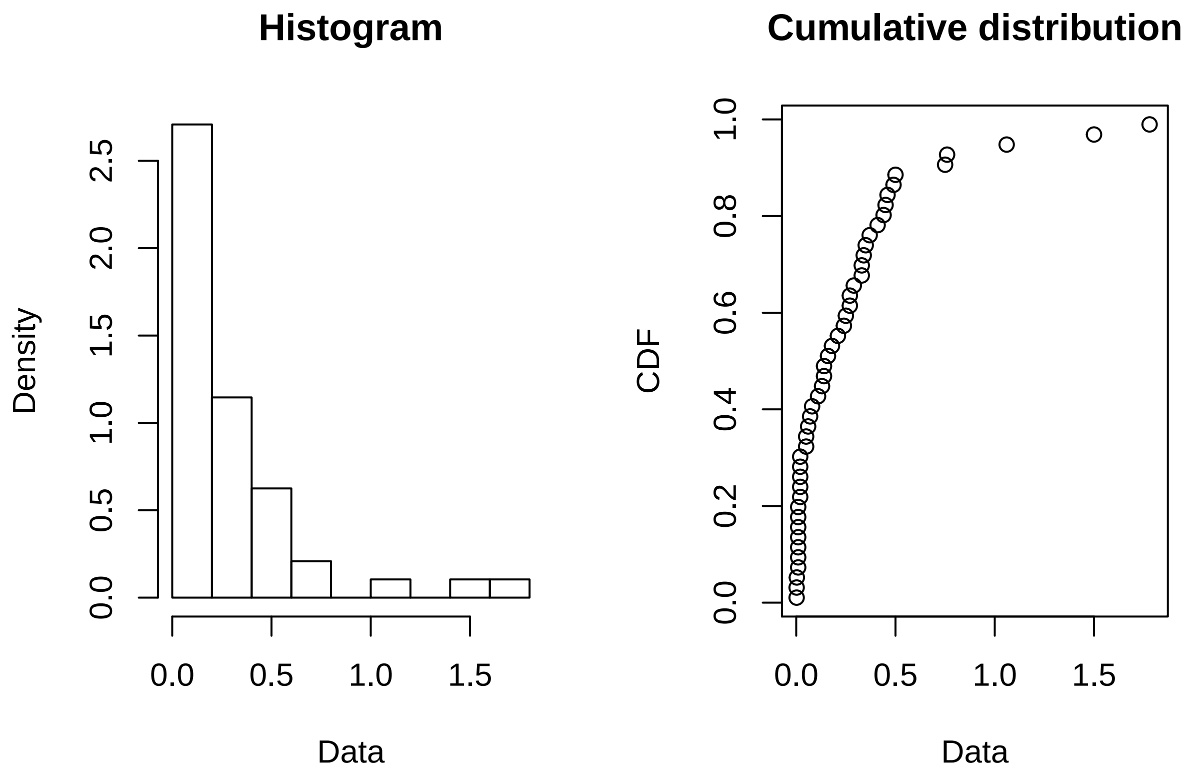
Rain in Southern Illinois

```
ill60 <- read.table("ill-60.txt")
year60<-as.numeric(as.array(ill60[,1]))
ill61 <- read.table("ill-61.txt")
year61<-as.numeric(as.array(ill61[,1]))
```

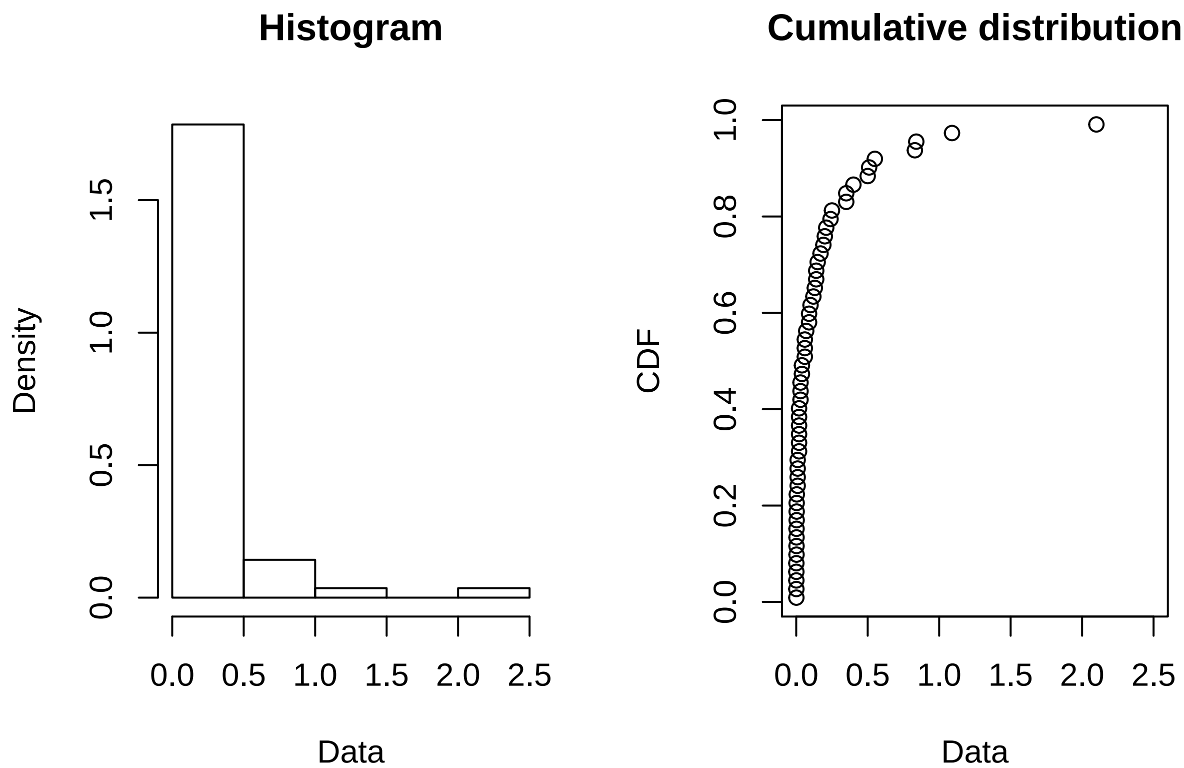
```
ill62 <- read.table("ill-62.txt")
year62<-as.numeric(as.array(ill62[,1]))
ill63 <- read.table("ill-63.txt")
year63<-as.numeric(as.array(ill63[,1]))
ill64 <- read.table("ill-64.txt")
year64<-as.numeric(as.array(ill64[,1]))
plotdist(year60)
```



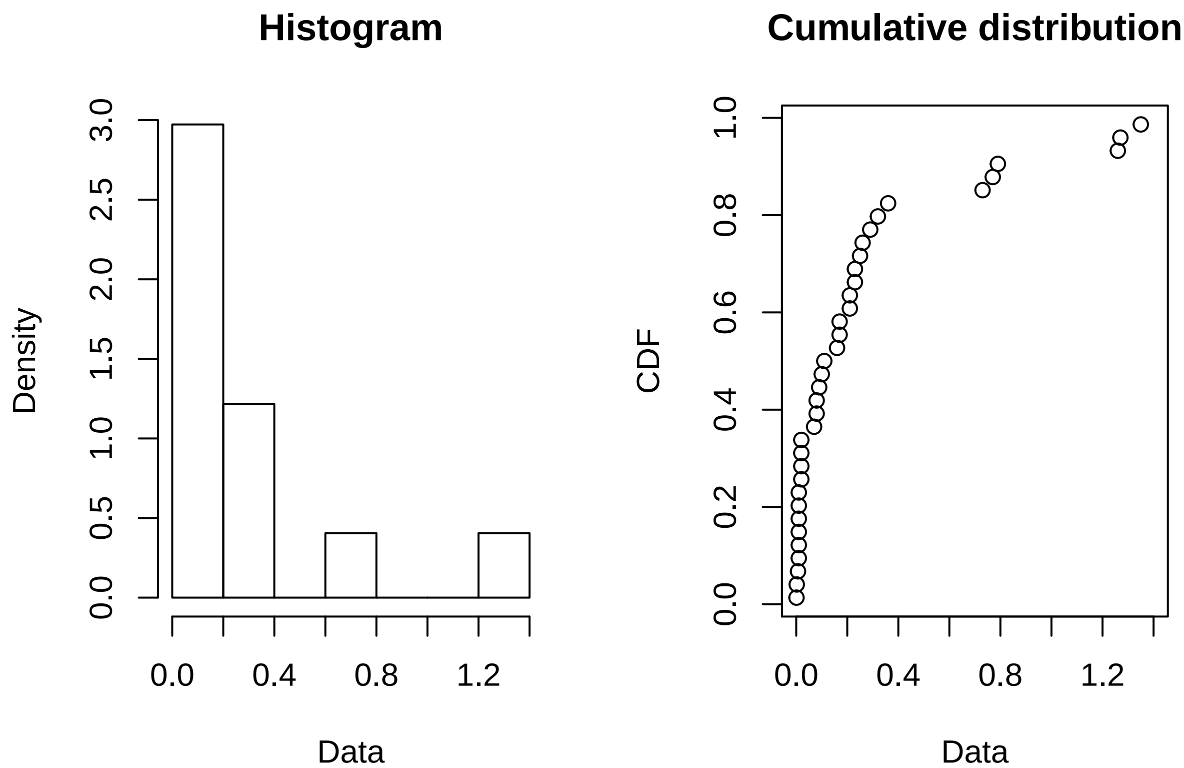

```
plotdist(year61)
```



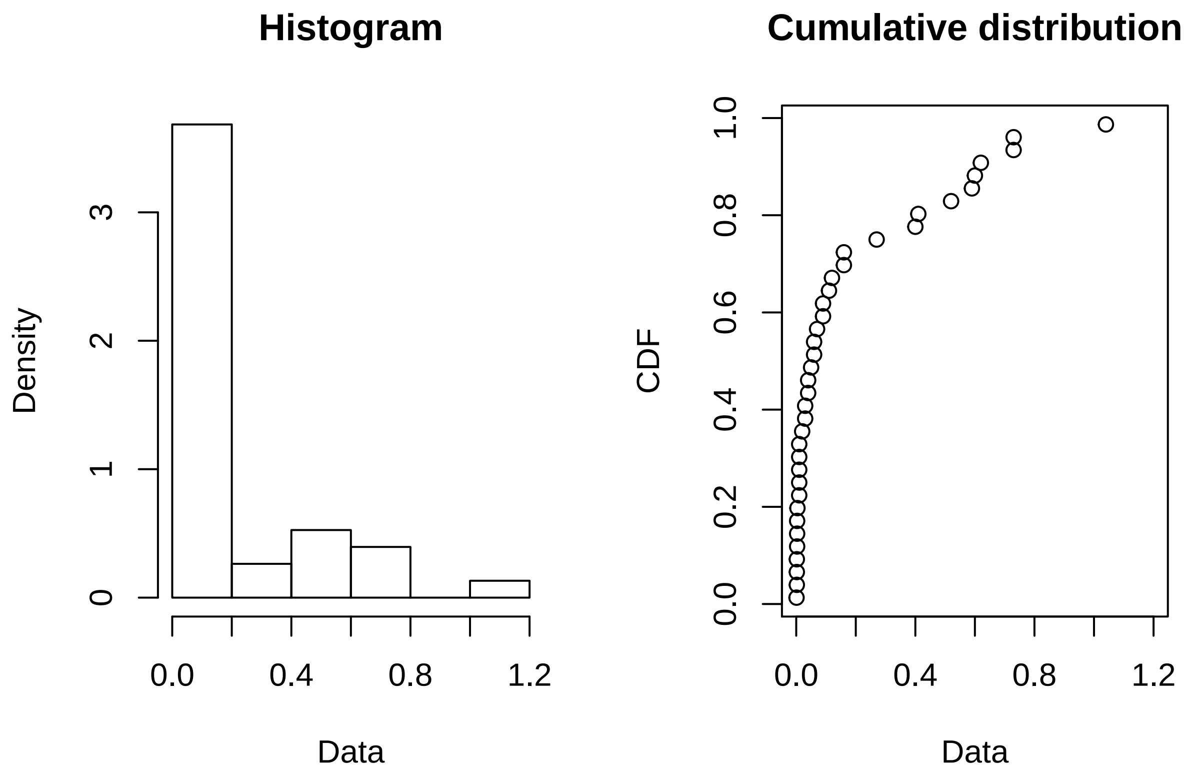

```
plotdist(year62)
```




```
plotdist(year63)
```




```
plotdist(year64)
```

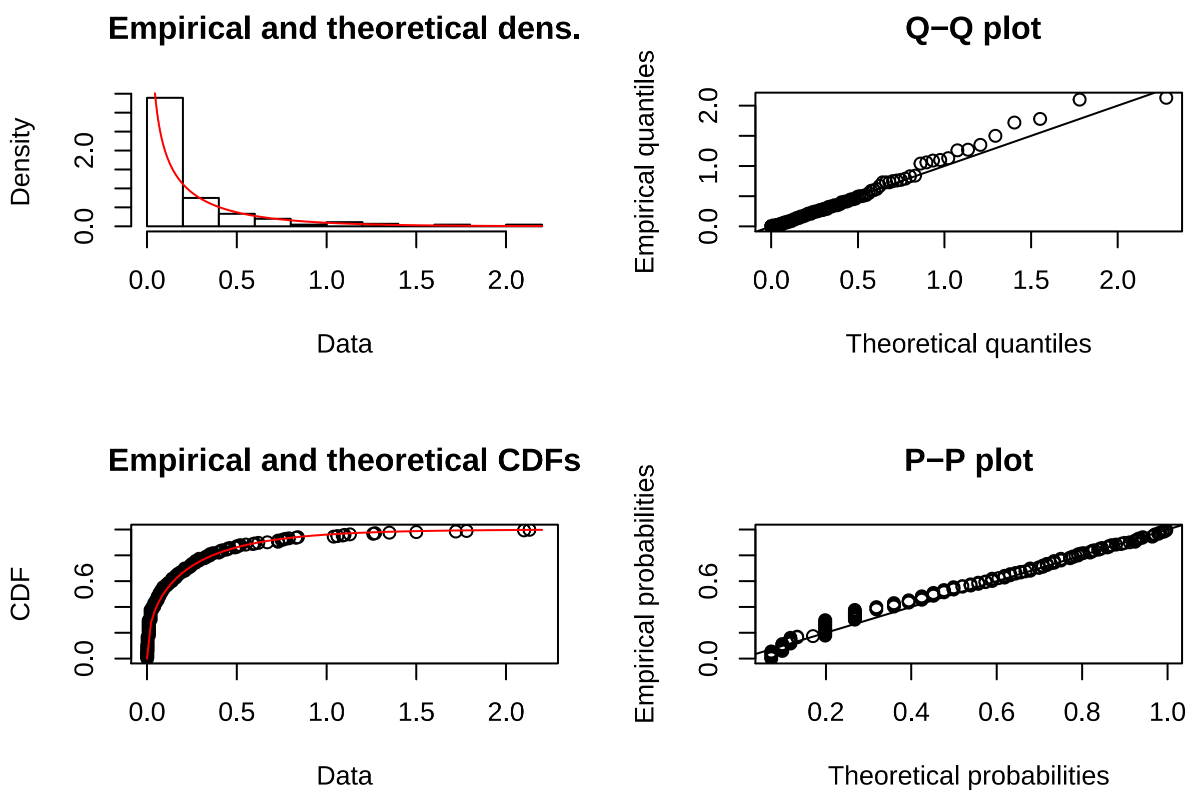



```
year <- c(1960,1961,1962,1963,1964)
total <- c(sum(year60),sum(year61),sum(year62),sum(year63),sum(year64))
sum <- as.data.frame(cbind(year,total))
kable(sum)
```

year	total
1960	10.574
1961	13.197
1962	10.346
1963	9.710
1964	7.110

Year 1961 is wetter based on the calculation for total amount of rainfall. However I cannot find any obvious different in those five distributions. Most of the rainfall are concentrated at the left side of data.

```
years <- c(year60,year61,year62,year63,year64)
gammadist <- fitdist(years, "gamma")
plot(gammadist)
```




```
summary(gammadist)
```

```
## 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
```

From the output, we can see that it fits well. Changnon and Huff are right about using gamma distribution.

```
gamma2 <- fitdist(years, "gamma",method = "mme")
gamma2d <- bootdist(gamma2)
summary(gamma2d)
```

```
## Parametric bootstrap medians and 95% percentile CI
##         Median      2.5%     97.5%
## shape 0.389167 0.2727101 0.5385677
## rate  1.739297 1.1448386 2.5462502
```

```
gamma3 <- fitdist(years, "gamma",method = "mle")
gamma3d <- bootdist(gamma3)
summary(gamma3d)
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4433395 0.3810743 0.5247192
## rate  1.9920329 1.5454244 2.5511166
```

Compare those two methods, mle has narrower CI. Therefore, I would choose mle as the estimator because it has the lower variance.

Decision theory