Final project

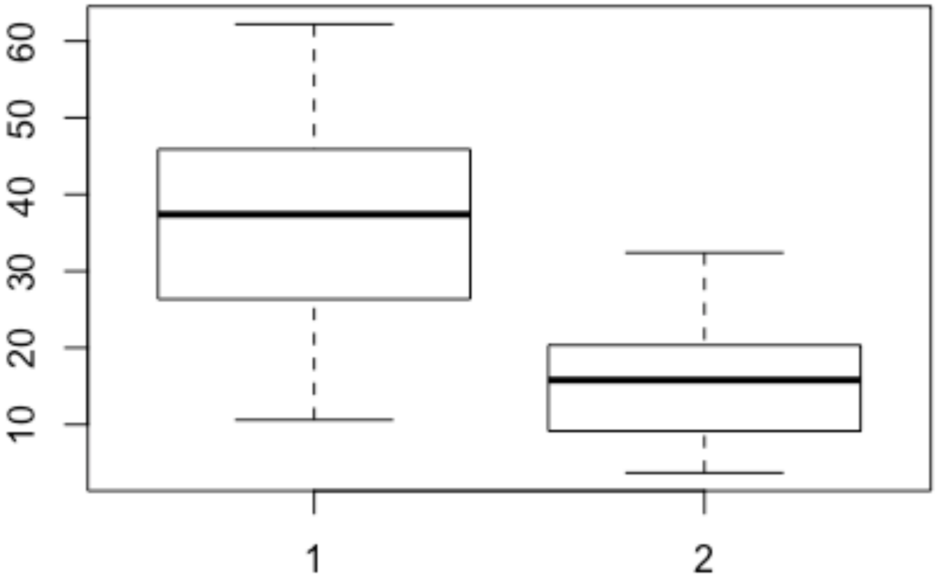
Qixuan Zhang

05/07/2019

1.Statistics and the Law

Initially, we propose the hypothesis as following: H0: Home morgage refusal rate of white applicants are the same as that of minority applicant H1: Home morgage refusal rate of white applicants are lower than that of minority applicant Next, we assume that the alpha level is 0.05 and we construct a two-sample t-test to test the argument if there is sufficient evidence for discrimination. Then we will do another two-sample t-test to test if there is sufficient evidence for discriminatiion bewteen high income white and high income minority.

```r
#Consturct the assumption
acorn<-read.csv("/Users/qixuanzhang/Desktop/acorn.csv")
boxplot(acorn$MIN,acorn$WHITE)
```



```
res.ftest <- var.test(acorn$MIN,acorn$WHITE)
res.ftest
```

```
##
##  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
```

The p-value of F-test is p = 0.02993. It's elss than the significance level alpha = 0.05. In conclusion, there is significant difference between the variances of the two sets of data. Therefore, we can use t-test witch assume unequal variances.

```
#T-test
res <- t.test(acorn$MIN, acorn$WHITE, var.equal = FALSE,alternative = "greate
r")
res
```

```
##
##  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
```

Conclusion:

The p-value of the test is 2.979e-07, which is less than the significance level alpha = 0.05. We reject the null hypothesis and support the argument that the data are sufficient evidence of discrimination to warrant corrective action.

2. Comparing suppliers

To compare the quality of ornithopters among three high schools: a) Area 51 Regional High b) BDV American Borstal c) Giffen Prep, we will use chi-square test to conduct the analysis since the quality of ornithopters(response variable) are categorical, instead of continuous one.

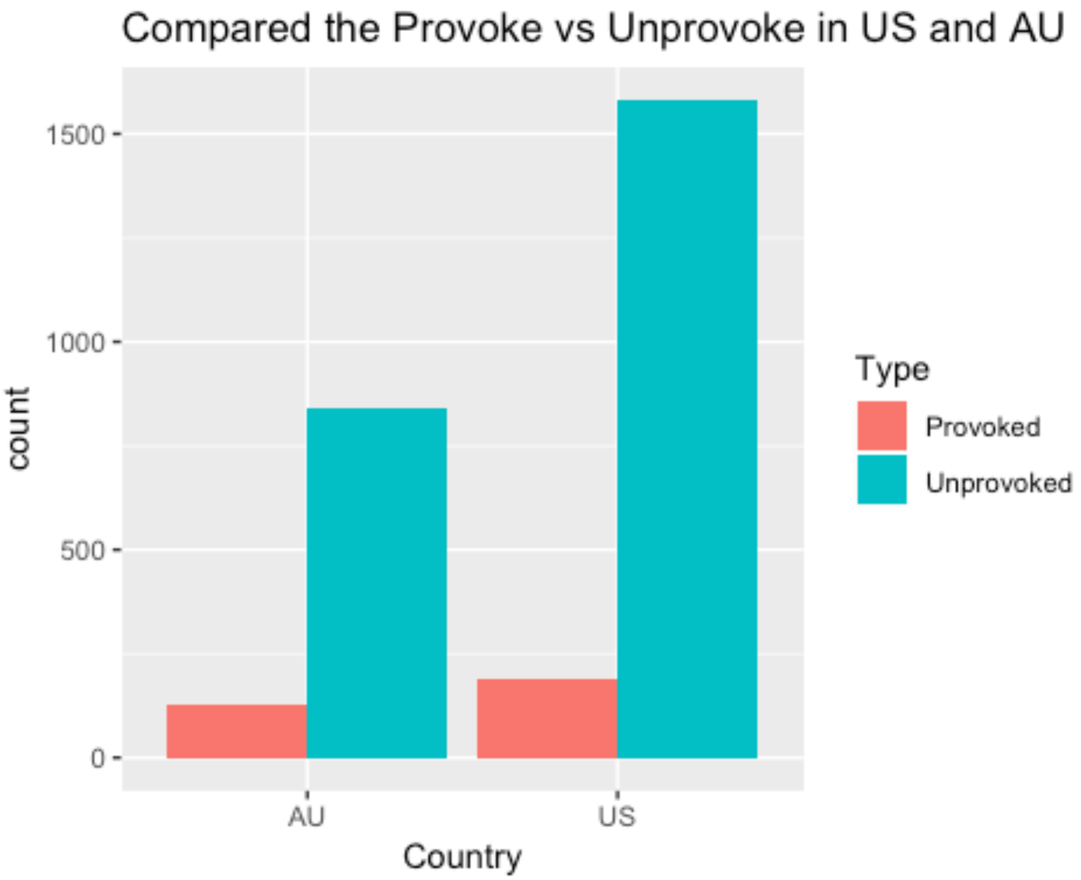
Ho: The ornithopters made by three schools produce the same qualityies Ha: The ornithopters made by all three shcools produce different qualities After that we did the chi-square test as following:

```
# Create a matrix
orn <- matrix(c(12,23,89,8,12,62,21,30,119),ncol=3,nrow = 3,byrow=TRUE)
colnames(orn) <- c("dead","art","fly")
rownames(orn) <- c("Area51","BDV","Giffen")
orn <- as.table(orn)
chisq.test(orn,correct = F)

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

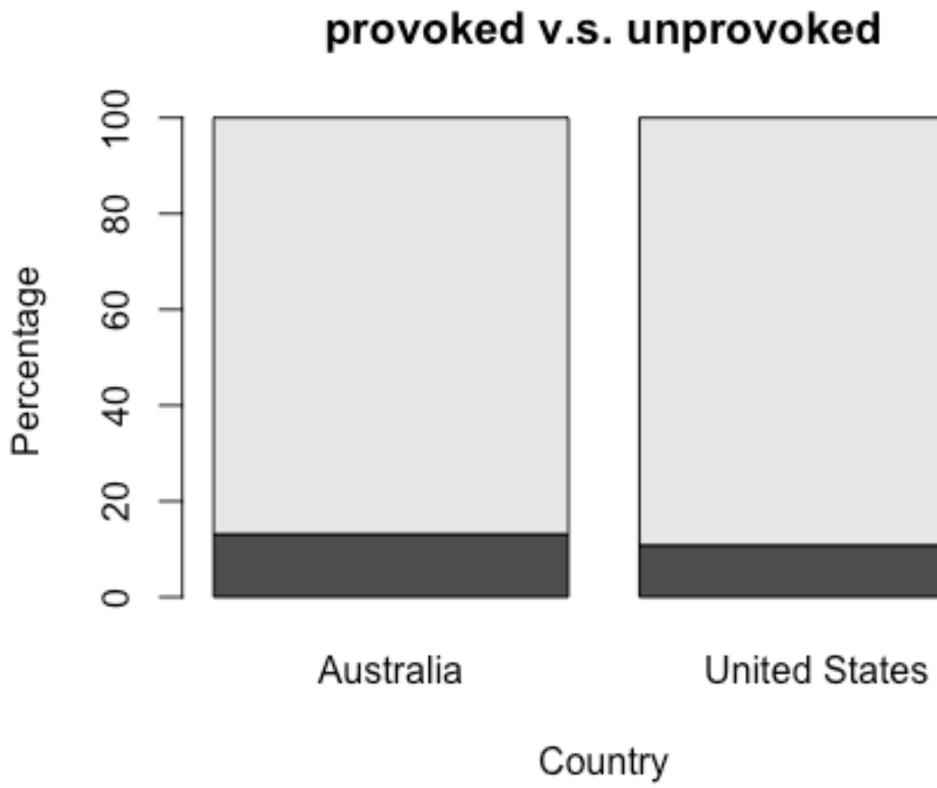
Based the result of the chi-square test, we could get the following conclusion: Because the p-value is 0.86 greater than 0.05, we fail to reject the null hypothesis and we have sufficient evidence to show that ornithopters made by all three schools produce the same qualities.

3. How deadly are sharks

```
shark<-read.csv("/Users/qixuanzhang/Desktop/sharkattack.csv")
head(shark)

##   X      Date       Country Country.code       Type     Continent
## 1 0 2/11/2017 United States           US Unprovoked North America
## 2 1  2/6/2017        Brazil           BR   Provoked South America
## 3 2  2/6/2017  South Africa           ZA       Boat        Africa
## 4 3  2/1/2017 United States           US       Boat North America
## 5 4  2/1/2017       Bahamas           BS Unprovoked North America
## 6 5 1/22/2017 United States           US Unprovoked North America
##   Hemisphere Activity Fatal
## 1          N Swimming     N
## 2          S    Other     N
## 3          S  Fishing     N
## 4          N    Other     N
## 5          N   Diving     N
## 6          N    Other     N

dead<-shark%>%
  filter(Country.code=="US"|Country.code=="AU")%>%
  filter(Type=="Provoked"|Type=="Unprovoked") %>%
  filter(Fatal=="N"|Fatal=="Y")

dead1<- dead %>%
  group_by(Type,Country.code)%>%
  summarise(count=n())
#View(dead1)
ggplot(dead1,aes(Country.code,count,fill=Type))+geom_bar(position="dodge", st
at="identity")+labs(title = "Compared the Provoke vs Unprovoke in US and AU",
x="Country")
```



```
dead.tbl<-table(droplevels(dead$Type),droplevels( dead$Country))
dead_percent<-apply(dead.tbl, 2, function(x){x*100/sum(x,na.rm=T)})
head(dead_percent)

## 
##              Australia United States
##   Provoked      13.147       10.8169
##   Unprovoked    86.853       89.1831

barplot(dead_percent,xlab="Country", ylab="Percentage",main = "provoked v.s.
unprovoked")
```

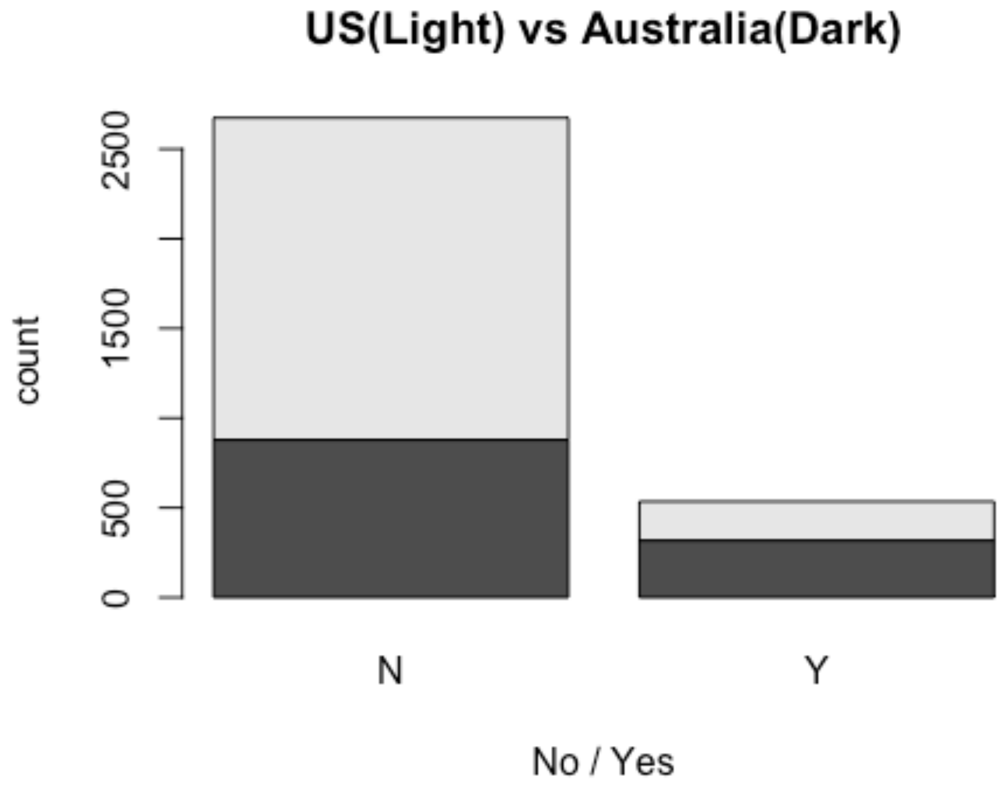


Just by looking at the provoked v.s. unprovoked in US and Australia (provoked in green), They are at about the same proportion; that is, provoked take 15% of all shark attacks in Australia and it takes 12% of all shark attacks in US.

```
fatal<-dead<-shark%>%
  filter(Country.code=="US"|Country.code=="AU")%>%
  filter(Fatal=="Y"|Fatal=="N")
fatal<-table(droplevels(fatal$Country),droplevels(fatal$Fatal))

knitr::kable(fatal)
```

	N	Y
Australia	879	318
United States	1795	217

```
barplot(fatal, xlab="No / Yes",ylab="count", main = "US(Light) vs Australia(D
ark)")
```



We can also look at the fatal v.s non-fatal sharks attacks in these two countries. From the table, we can see that there are definitely more sharks attacks in US. However, the proportion of fatal attacks in Australia (26.5%) is way higher than that in US (10%). To futuer test that Sharks Australia are more deadly, or fatal than those in US, we conduct a chi-square test.

```
# Applied chi-square test
chisq.test(fatal,correct = F)

##
##  Pearson's Chi-squared test
##
## data:  fatal
## X-squared = 134.54, df = 1, p-value < 2.2e-16

prob <- matrix(c(0.2739171,0.5593643,0.09909629,0.06762231), nrow=2,
               dimnames = list(c("Australi","US"),c("NonFatal","Fatal")))
prob

##            NonFatal      Fatal
## Australi 0.2739171 0.09909629
## US       0.5593643 0.06762231
```

```
io<-879+318+1795+217
pwr.chisq.test(w = ES.w2(prob), N = io, 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
```

From the chi-square test, we have sufficient evidence to say that that sharks attacks in Australia is more fatal, more deadly than sharks attacks in U.S, although the number of attacks in US is higher than that in Australia.With sample size equal to 2109, the statistical power of the test chi-square test is 1.

4. Power Analysis

Just like it is described in the book, the power to detect the difference between hypothetical parameters .65 and .45 is .48 while the power to detect the difference between hypothetical parameters .25 and .05 is .82, even though the difference between both pairs of values is .20, which means hypothetical parameters of this binomial distribution doesn not provide a scale of equal units of detectability because 0.25 and 0.05 fall into one extreme of the range.

However, after arcsine transformation, which transforms the proportional parameter (from 0 to 1) to the scale of $-\pi/2$ to $\pi/2$. and then transformed t1 -t2 = h, which has euqal dectectability. This can solve the problem of falling into either side of the range.

5. MLE

Case1 MLE of Exponential Distribution

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

$$L(\lambda; X_1, \ldots, X_n) = \lambda e^{-\lambda X_1} \lambda e^{-\lambda X_2} \ldots \lambda e^{-\lambda X_n}$$

$$L(\lambda; X_1, \ldots, X_n) = \lambda^n e^{-\lambda \sum X_i}$$

$$l(\lambda; X_1, \ldots, X_n) = n log(\lambda) - \lambda \sum X_i$$

$$\frac{dl(\lambda; X1, \ldots, Xn)}{d\lambda} = \frac{n}{\lambda} - \sum X = 0$$

$$\hat{\lambda} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}_n}$$

Case2 Moment Estimator and MLE for new distribution θ MOM:

$$\begin{aligned}
E[X] &= \int_0^1 x((1-\theta) + 2\theta x)dx \\
&= (1-\theta)\int_0^1 xdx + \int_0^1 2\theta x^2 dx \\
&= (1-\theta)\frac{1}{2}x^2 \mid_0^1 + 2\theta\frac{1}{3}x^3 \mid_0^1 \\
&= \frac{1}{2} - \frac{1}{2}\theta + \frac{2}{3}\theta \\
&= \frac{1}{6}\theta + \frac{1}{2}
\end{aligned}$$

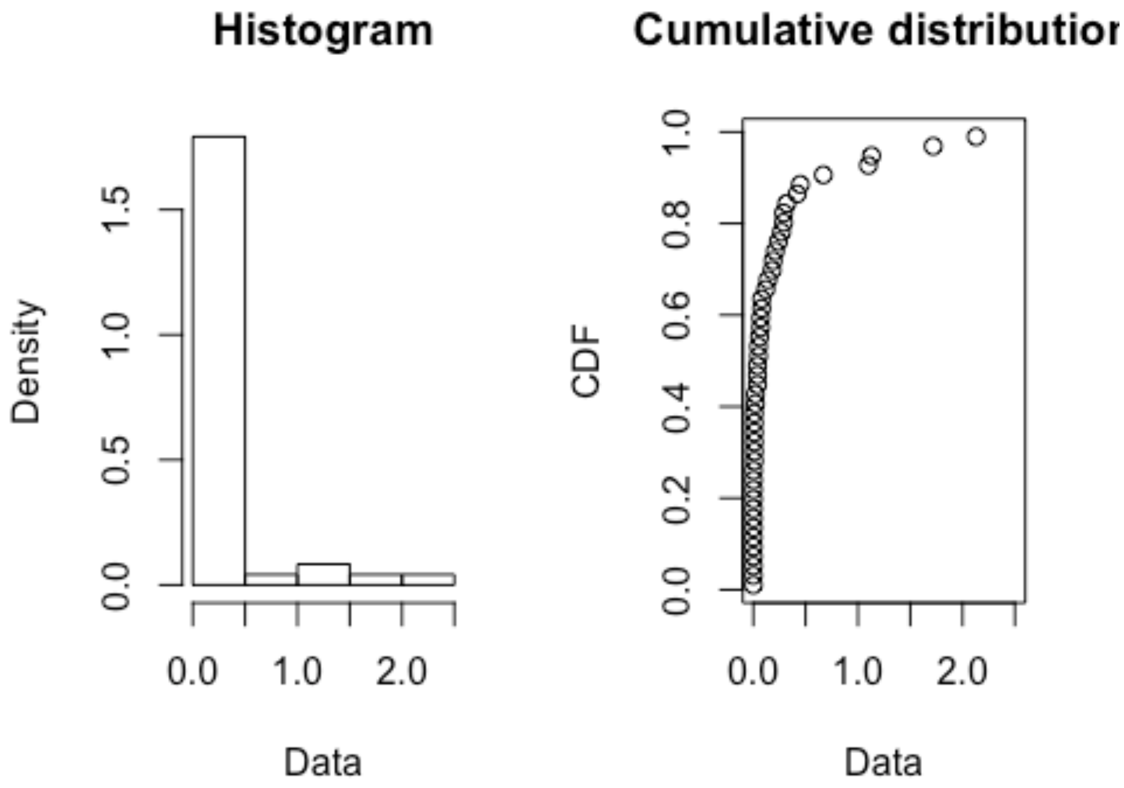
MLE:

$$L(\theta; X_1, \ldots, X_n) = [(1-\theta) + 2\theta X_1] \ldots [(1-\theta) + 2\theta X_n]$$

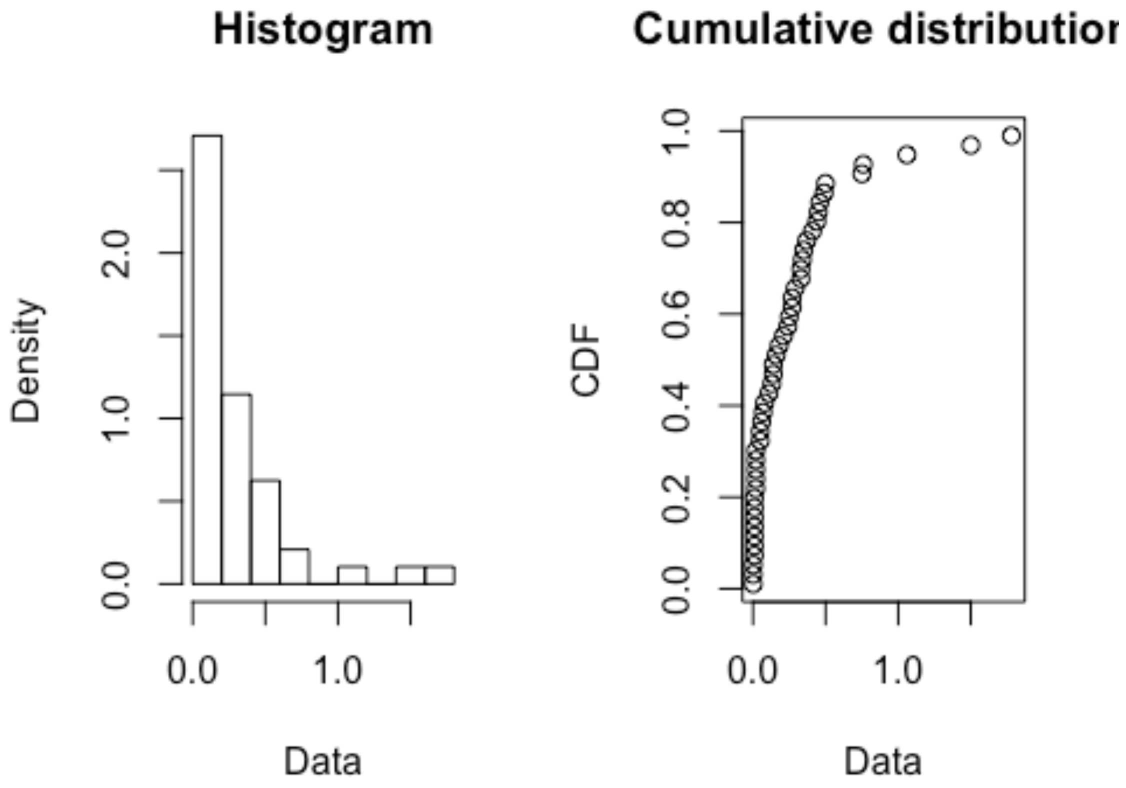
$$l(\theta; X_1, \ldots, X_n) = log[(1-\theta) + 2\theta X_1] \ldots log[(1-\theta) + 2\theta X_1]$$

6. Rain in Southern Illinois

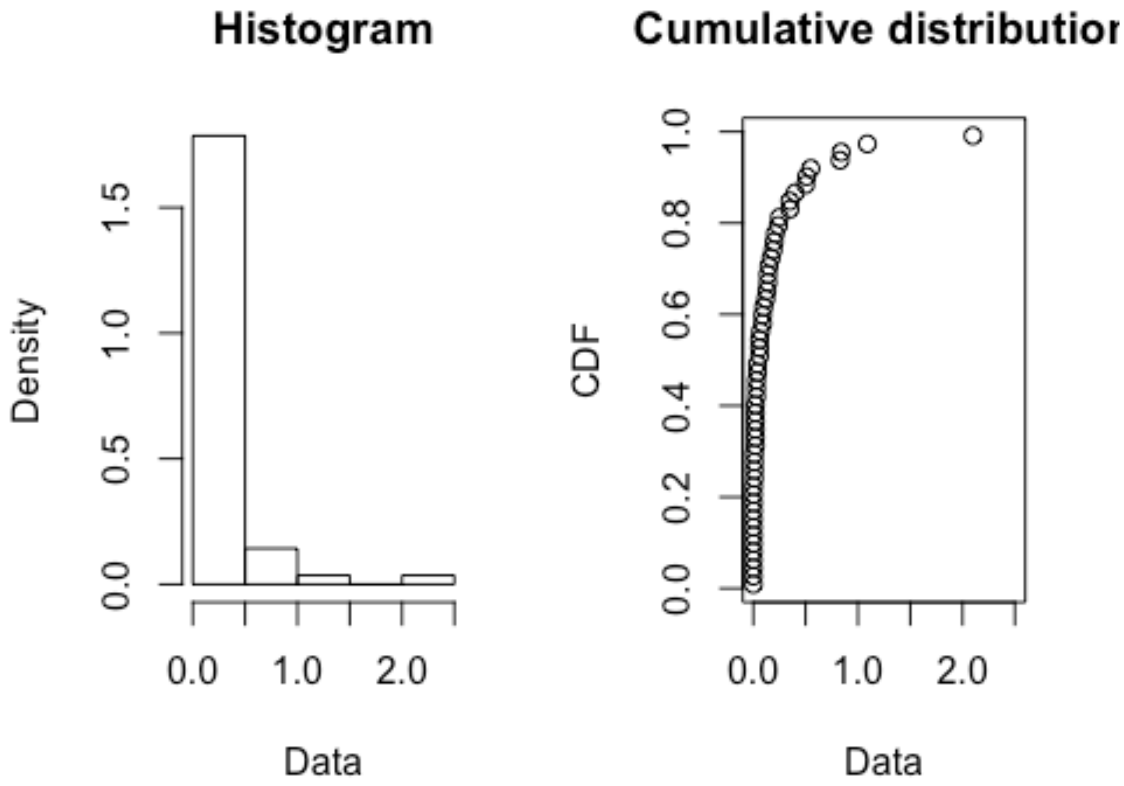
```
ill.60 <- read.table("~/Desktop/illinois storms/ill-60.txt", quote="\"", comm
ent.char="")
yr60<-as.numeric(as.array(ill.60[,1]))
ill.61 <- read.table("~/Desktop/illinois storms/ill-61.txt", quote="\"", comm
ent.char="")
yr61<-as.numeric(as.array(ill.61[,1]))
ill.62 <- read.table("~/Desktop/illinois storms/ill-62.txt", quote="\"", comm
ent.char="")
yr62<-as.numeric(as.array(ill.62[,1]))
ill.63 <- read.table("~/Desktop/illinois storms/ill-63.txt", quote="\"", comm
ent.char="")
yr63<-as.numeric(as.array(ill.63[,1]))
ill.64 <- read.table("~/Desktop/illinois storms/ill-64.txt", quote="\"", comm
ent.char="")
yr64<-as.numeric(as.array(ill.64[,1]))

library(fitdistrplus)

## Loading required package: MASS

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

## Loading required package: survival

## Loading required package: npsurv

## Loading required package: lsei

plotdist(yr60)
```



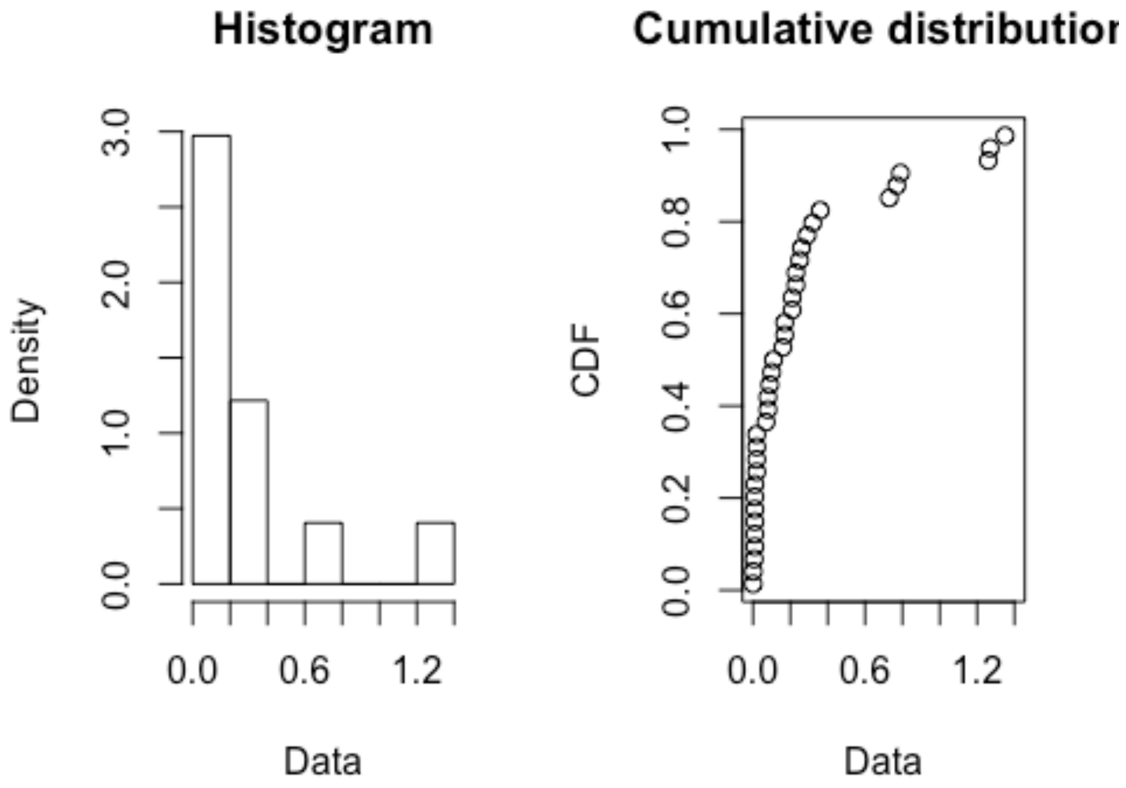
```
plotdist(yr61)
```



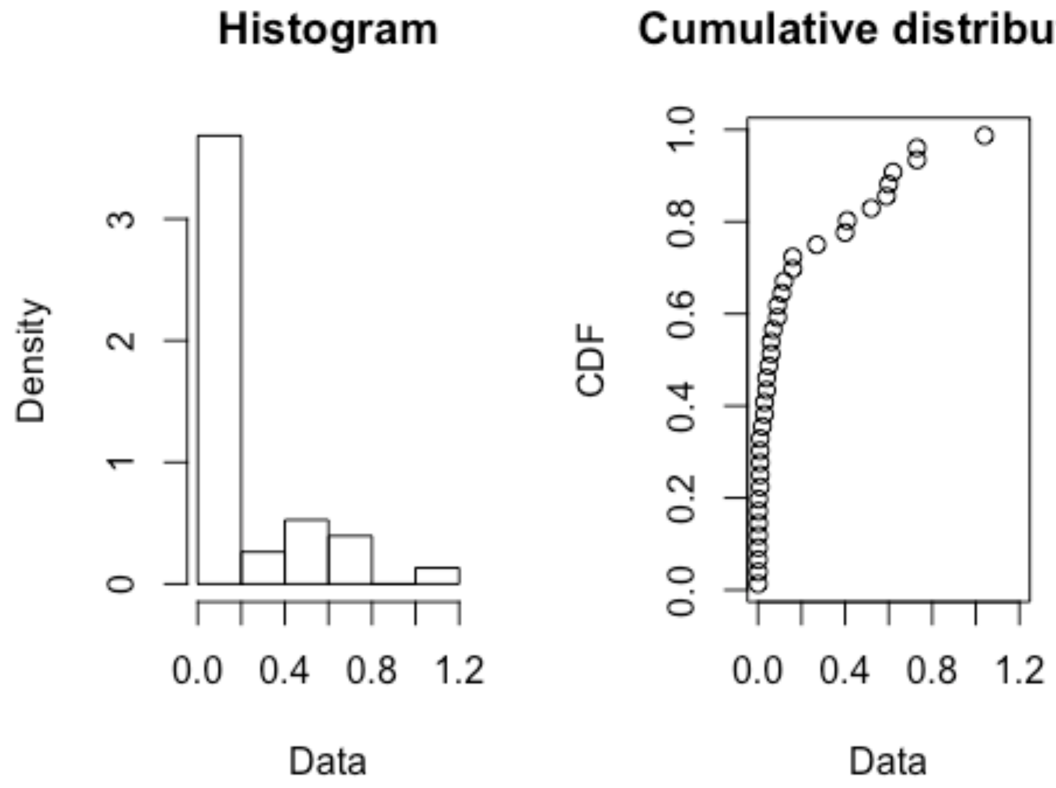
```
plotdist(yr62)
```



```r
plotdist(yr63)
```



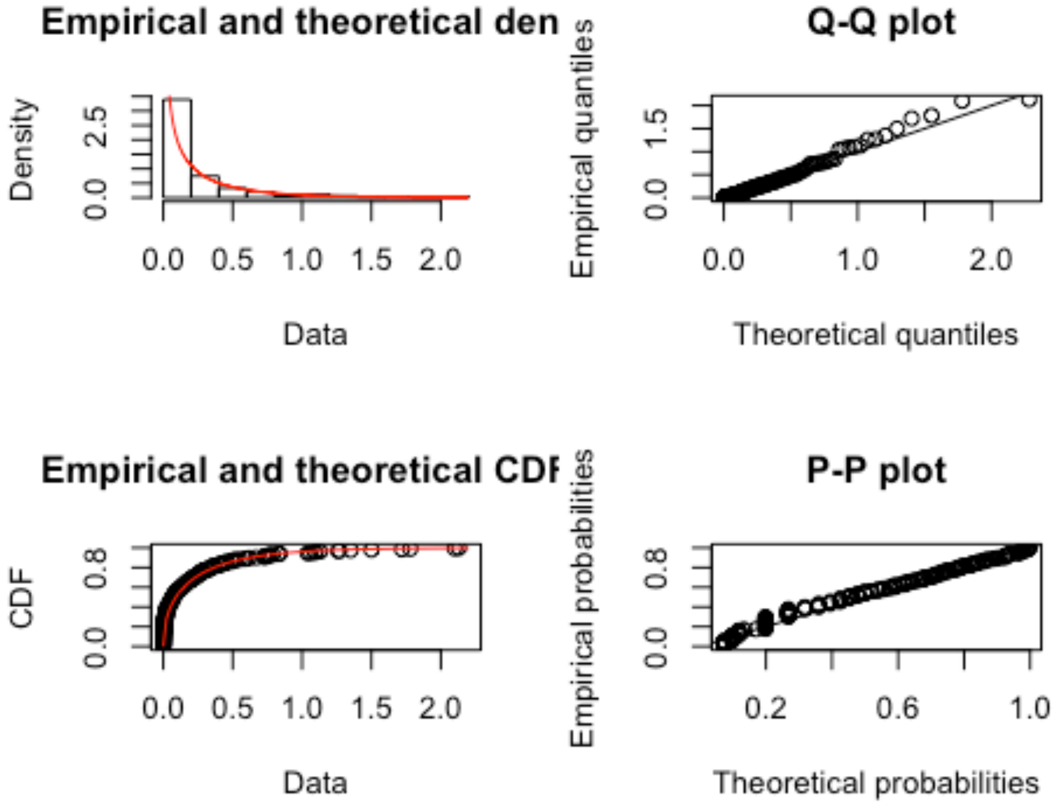
```
plotdist(yr64)
```



```
paste("The total rainfall for 1960 is",sum(yr60),sep = " ")

## [1] "The total rainfall for 1960 is 10.574"

paste("The total rainfall for 1960 is",sum(yr61),sep = " ")

## [1] "The total rainfall for 1960 is 13.197"

paste("The total rainfall for 1960 is",sum(yr62),sep = " ")

## [1] "The total rainfall for 1960 is 10.346"

paste("The total rainfall for 1960 is",sum(yr63),sep = " ")

## [1] "The total rainfall for 1960 is 9.71"

paste("The total rainfall for 1960 is",sum(yr64),sep = " ")

## [1] "The total rainfall for 1960 is 7.11"
```

From the total rainfall of each year, the rainfall in 1961 is definitely wetter because of 13.2 units of rainfall in total. From the distribution of rainfall in year 1960, we know that this year is wetter because of storms, though not too many, produce more rain each time compared to rainfall in other 4 years. In terms of all 5 distribution in these years, they are similar because most of rainfall among are concentrated on the left side of the distribution.

```r
#Fit Gamma Distribution
all_rainfall<-c(yr60,yr61,yr62,yr63,yr64)
gamma <- fitdist(all_rainfall, "gamma")
plot(gamma)
```



```
summary(gamma)

## 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
```

Using the fitdist() funtion, we fit the gamma distribution into all observed rainfall data of four year from 1960 to 1964. From the density and QQplots, the gamma distribution fits really well on this data. Changnon and Huff are definitely right about using gamma distribution.

```
fmm <- fitdist(all_rainfall, "gamma",method = "mme")
bmm <- bootdist(fmm)
summary(bmm)
```

```
## Parametric bootstrap medians and 95% percentile CI
##           Median      2.5%     97.5%
## shape 0.3968997 0.2764507 0.5228124
## rate  1.7679591 1.1714927 2.4691954
```

```
fgamma <- fitdist(all_rainfall, "gamma",method = "mle")
bgmm <- bootdist(fgamma)
summary(bgmm)
```

```
## Parametric bootstrap medians and 95% percentile CI
##           Median      2.5%     97.5%
## shape 0.4447916 0.3818173 0.5163182
## rate  1.9919889 1.5394412 2.5860336
```

For method of moment the 95% confidence interval of shape from bootstrap sample is (0.28,0.53), the rate is (1.17,2.62). For MLE, the 95% confidence interval of shape from bootstrap sample is (0.38,0.51),the rate is (1.57,2.59). Apparently, the MLE estimates have narraow CI and thus lower variances.I would choose to present MLE as the estimator because it has lower variance.

7. Decision Theory

$$P(x_1,\dots,x_n|\delta) = \prod_{i=1}^{N} P\,(X = x_i|\delta) = \prod_{i=1}^{N} \delta^{x_i}\,(1-\delta)^{x_i}$$

Let n be number of β with value 1

$$(\beta_s, s \in S) = (0,1)$$

$$P(x_1,\dots,x_n|\delta) = \delta^n(1-\delta)^{N-n}$$

Prior:

$$P(\delta) = Beta(c,d) = \frac{1}{B(c,d)}\delta^{c-1}(1-\delta)^{d-1}$$

From Bayes

$$P(\delta|x_1,\dots,x_n) = P(x_1,\dots,x_n|\delta)P(\delta) = \frac{1}{C}\delta^{c+n-1}(1-\delta)^{N-n+d-1}$$

Where $C = \int P(x_1,\dots,x_n,\delta)d\delta$ is the normalizing constant. But notice that this posterior is also a Beta distribution, with new parameters:

$$\delta|x_1,\dots,x_n \sim Beta(c+n, d+N-n)$$

$$\beta = E(\delta | x_1, \ldots, x_n) = \frac{c + n}{c + d + N}$$

From:

$$\delta(n) = 0 \quad for \quad \beta < \alpha$$
$$\delta(n) = \lambda \quad for \quad \beta = \alpha, \quad where \quad 0 < \lambda < 1$$
$$\delta(n) = 1 \quad for \quad \beta > \alpha$$

We have:

$$\delta(n) = 0 \quad for \quad (c + n)/(c + d + N) < \alpha$$
$$\delta(n) = \lambda \quad for \quad (c + n)/(c + d + N) = \alpha, \quad where \quad 0 < \lambda < 1$$
$$\delta(n) = 1 \quad for \quad (c + n)/(c + d + N) > \alpha$$

We use the following code to reproduce the table:

```
library(data.table)

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following objects are masked from 'package:reshape2':
##
##     dcast, melt

library(tidyverse)

## ── Attaching packages ────────────────────────────────── tidyverse 1.2.1 ─
─

## ✔ tibble  1.4.2     ✔ purrr   0.2.5
## ✔ tidyr   0.8.2     ✔ stringr 1.3.1
## ✔ readr   1.1.1     ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────── tidyverse_conflicts() ─
─

## ✖ data.table::between() masks dplyr::between()
## ✖ dplyr::filter()       masks stats::filter()
## ✖ data.table::first()   masks dplyr::first()
## ✖ dplyr::lag()          masks stats::lag()
## ✖ data.table::last()    masks dplyr::last()
## ✖ car::recode()         masks dplyr::recode()
## ✖ MASS::select()        masks dplyr::select()
## ✖ purrr::some()         masks car::some()
## ✖ purrr::transpose()    masks data.table::transpose()

# get wide and long format of table 1
table1 <- fread("table1.csv",skip = 2, nrows = 5)
table1[,"alpha"] <-c(0.1,0.25,0.5,.75,.9)
table1$V1<-NULL
colnames(table1)[1:11] <- 0:10
tbl1 <- gather(table1,"N","n0",-alpha)
# get wide and long format of table 2
table2 <- fread("table1.csv",skip = 8, nrows = 5)
table2[,"alpha"] <-c(0.1,0.25,0.5,.75,.9)
table2$V1<-NULL
colnames(table2)[1:11] <- 0:10
tbl2 <- gather(table2,"N","lambda",-alpha)
tbl<-left_join(tbl1,tbl2,by=c("alpha"="alpha","N"="N"))
tbl$N <- as.numeric(tbl$N)
tbl$n0 <- as.numeric(tbl$n0)
tbl$lambda <- as.numeric(tbl$lambda)

beta <- seq(0,1,0.01)
delta <- function(n0,lambda,n){
  if (n<n0){
    return(0)
  }
  else if (n==n0){
    return(lambda)
  }
  else {
    return(1)
  }
}
E <- function(n0,lambda,N){
  sum = c
  for (i in 0:N){
    f = factorial(N)/(factorial(i)*factorial(N-i))*beta^(i)*(1-beta)^(N-i)
    delta1 = delta(n0,lambda,i)
    sum = sum+ f*delta1
  }
  return(sum)
}
W <- c()
table_reproduce<- matrix(nrow = nrow(tbl),ncol = length(beta))
```

The top panel of the table shows that the threshold n0 of experimental successes for allocation of persons to treatment B increases with the sample size and with the success probability of treatment A. The inequality |n0 – αN|≤≤ 1 holds everywhere in the table. Thus, the minimax-regret rule is well approximated by an empirical success rule. The third panel shows that the value of minimax regret decreases by roughly an order of magnitude as the sample size increases from to 10. For example, when α= 0.50, it falls from 0.25 to 0.027. Thus, even a sample size as small as 10 suffices to make maximum regret quite small.