MA 677 - Spring 2019, Final project

Jinfei Xue

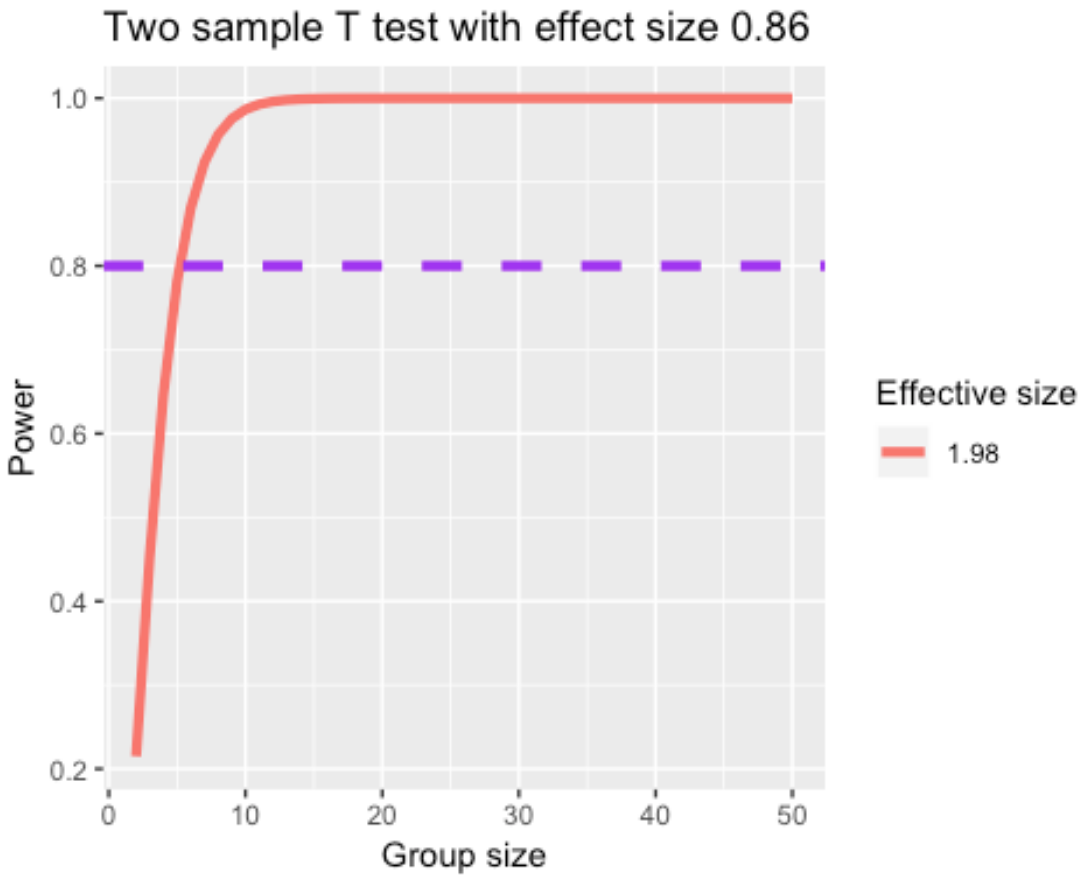
5/3/2019

```
library(tidyverse)
library(stringr)
library(gridExtra)
library(kableExtra)
library(ggpubr)
library(pwr)
library(fitdistrplus)
```

1. Statistics and the Law

```
#data
MIN<-
c(20.90,23.23,23.10,30.40,42.70,62.20,39.5,38.40,26.20,55.90,49.70,44.60,36.40,32.00,10.60,34.30,42.30,26.50,51.50,47.20)
WHITE<-
c(3.7,5.5,6.7,9.0,13.9,20.6,13.4,13.2,9.3,21.0,20.1,19.1,16.0,16.0,5.6,18.4,23.3,15.6,32.4,29.7)
acorn <- data.frame(MIN, WHITE)

#Calculate effect size
cohens_d <- function(x, y) {
lx <- length(x) - 1
ly <- length(y) - 1
md <- abs(mean(x) - mean(y)) ## mean difference (numerator)
csd <- lx * var(x) + ly * var(y)
csd <- csd / (lx + ly)
csd <- sqrt(csd) ## common sd computation
cd <- md / csd ## cohen's d
return(cd)
}
effect_size <- cohens_d(acorn$MIN, acorn$WHITE)
pwr.t.test(
n = dim(acorn)[1], effect_size,
sig.level = 0.05, power = NULL, type = c("two.sample")
)

##
##      Two-sample t test power calculation
##
##               n = 20
##               d = 1.977454
```

```
##       sig.level = 0.05
##           power = 0.9999818
##     alternative = two.sided
##
## NOTE: n is number in *each* group

pwr.t.test(
n = NULL, effect_size,
sig.level = 0.05, power = 0.95, type = c("two.sample")
)

##
##      Two-sample t test power calculation
##
##               n = 7.754825
##               d = 1.977454
##       sig.level = 0.05
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group

# Plot effect size
# Reference: Ben's Consulting Group 4
n <- seq(2, 50, by = 1)
plot_effectsize <- function(n, effect_size) {
ptab1 <- cbind(NULL)
for (i in seq(2, 50, by = 1)) {
pwrt1 <- pwr.t2n.test(
n1 = i, n2 = i,
sig.level = 0.05, power = NULL,
d = effect_size, alternative = "two.sided"
)
ptab1 <- rbind(ptab1, pwrt1$power)
}
temp <- as.data.frame(ptab1)
colnames(temp)[1] <- "num"
ggplot(temp) +
geom_line(aes(x = n, y = num, colour = "darkblue"), size = 1.5) +
scale_color_discrete(name = "Effective size", labels = c(round(effect_size,
2))) +
geom_hline(yintercept = 0.8, linetype = "dashed", color = "purple", size =
1.5) +
ylab("Power") + scale_y_continuous(breaks = seq(0, 1, by = 0.2)) +
ggtitle("Two sample T test with effect size 0.86") + xlab("Group size")
}
plot_effectsize(n, effect_size)
```




```
# Perform t-test
t.test(acorn$MIN, acorn$WHITE)

##
##  Welch Two Sample t-test
##
## data:  acorn$MIN and acorn$WHITE
## t = 6.2533, df = 31.028, p-value = 5.958e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  14.3239 28.1891
## sample estimates:
## mean of x mean of y
##   36.8815   15.6250
```

Through power analysis, we found that the power of this t test only has 0.9999818 which is not relatively high. If we want our power at least equal to 0.95, then we will only need almost 8 samples in each group which our sample size already way more than that. So the data are sufficient for us to perform t-test. The p-value of MIN vs WHITE t-tests is smaller than 0.05, which indicates that there is a discrimination between the rates of mortgage application refusals of minority applications and white applications.

2. Comparing Suppliers

H0:They all produce about the same quality H1:They do not produce about the same quality

```
data2 <- matrix(c(12,23,89,8,12,62,21,30,119),ncol=3,nrow = 3,byrow=TRUE)
colnames(data2) <- c("dead","art","fly")
rownames(data2) <- c("Area51","BDV","Giffen")
fly <- as.table(data2)
chisq.test(data2,correct = F)

##
##  Pearson's Chi-squared test
##
## data:  data2
## X-squared = 1.3006, df = 4, p-value = 0.8613
```

The p-value of this chi-square test is 0.8613, which is much greater than the significant level alpha=0.05. Therefore, we fail to reject the null hypothesis. The data are sufficient to show that three schools produce the same quality.

3. How deadly are sharks?

```
# read data
sharkattack <- read.csv("sharkattack.csv")

# Filter US and AU shark attack records out from the original dataset
us_shark <- sharkattack[which(sharkattack$Country.code == "US"), ]
au_shark <- sharkattack[which(sharkattack$Country.code == "AU"), ]
# Drop Unknown from Fatal
us_shark <- us_shark[which(us_shark$Fatal != "UNKNOWN"), ]
au_shark <- au_shark[which(au_shark$Fatal != "UNKNOWN"), ]
# Create binary variable for Fatal
us_shark$Fatal.code <- ifelse(us_shark$Fatal == "Y", 1, 0)
au_shark$Fatal.code <- ifelse(au_shark$Fatal == "Y", 1, 0)
# Check the size of two samples
dim(us_shark)[1] == dim(au_shark)[1]

## [1] FALSE

# Calculate effect size
# Reference: https://stackoverflow.com/questions/15436702/estimate-cohens-d-
for-effect-size
cohens_d <- function(x, y) {
  lx <- length(x) - 1
  ly <- length(y) - 1
  md <- abs(mean(x) - mean(y)) ## mean difference (numerator)
  csd <- lx * var(x) + ly * var(y)
  csd <- csd / (lx + ly)
  csd <- sqrt(csd) ## common sd computation
```

```r
  cd <- md / csd ## cohen's d
}
cohens_d(au_shark$Fatal.code, us_shark$Fatal.code)
# Alternative way to calculate effect size
ES.h(mean(au_shark$Fatal.code), mean(us_shark$Fatal.code))
```

```
## [1] 0.4137712
```

```r
# Compute power of test
pwr.2p2n.test(
h = 0.4324294, n1 = dim(au_shark)[1],
n2 = dim(us_shark)[1], sig.level = 0.05,
alternative = "greater"
)
```

```
##
##      difference of proportion power calculation for binomial distribution
(arcsine transformation)
##
##               h = 0.4324294
##              n1 = 1197
##              n2 = 2012
##       sig.level = 0.05
##           power = 1
##     alternative = greater
##
## NOTE: different sample sizes
```

```r
pwr.2p2n.test(
h = 0.4137712, n1 = dim(au_shark)[1],
n2 = dim(us_shark)[1], sig.level = 0.05,
alternative = "greater"
)
```

```
##
##      difference of proportion power calculation for binomial distribution
(arcsine transformation)
##
##               h = 0.4137712
##              n1 = 1197
##              n2 = 2012
##       sig.level = 0.05
##           power = 1
##     alternative = greater
##
## NOTE: different sample sizes
```

```r
# Two-proportions Z-test
prop.test(
x = c(
  sum(au_shark$Fatal.code == 1),
```

```
  sum(us_shark$Fatal.code == 1)
),
n = c(dim(au_shark)[1], dim(us_shark)[1]),
alternative = "greater"
)

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(sum(au_shark$Fatal.code == 1), sum(us_shark$Fatal.code == 1)) out
of c(dim(au_shark)[1], dim(us_shark)[1])
## X-squared = 133.41, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1332633 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.2656642 0.1078529
```

First of all, I used two different function to calculate the effect size. Then "pwr.2p2n.test" was formed for power analysis since the sample sizes of US and AU shark attack are different. Based on that, I got 2 results of power with two different effect sizes, both of them equal to 1, which is pretty high comparatively. So we could know that the result of proportion test is reliable. The p-value of two-proportions z-test with less than 2.2e-16 indicates that we should reject the null hypothesis: the proportion of two samples are equal. In conclusion, the sharks in Australia were, on average, a more vicious lot than the sharks in the United States.

4. Power analysis

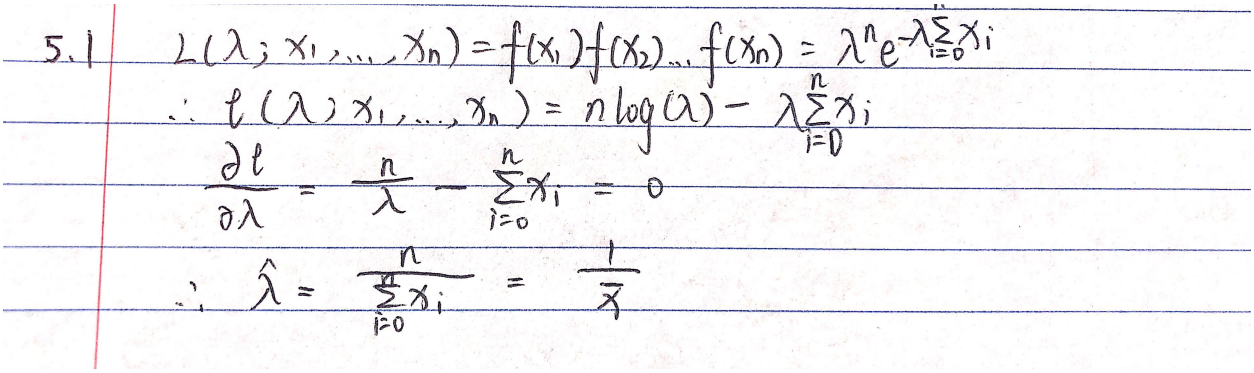
Arcsine transformation severs for the problem that P does not provide a scale of equal units of detectability. It uses an non-linear transformation on P so that after arcsine transforming of P, equal differences between units are equally detectable. The differences between ES index gives values whose delectability does not depend on whether the transformation of P or P itself fall around the middle or on one side of their possible range.

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. Estimators

5.1 Exponential



5.1

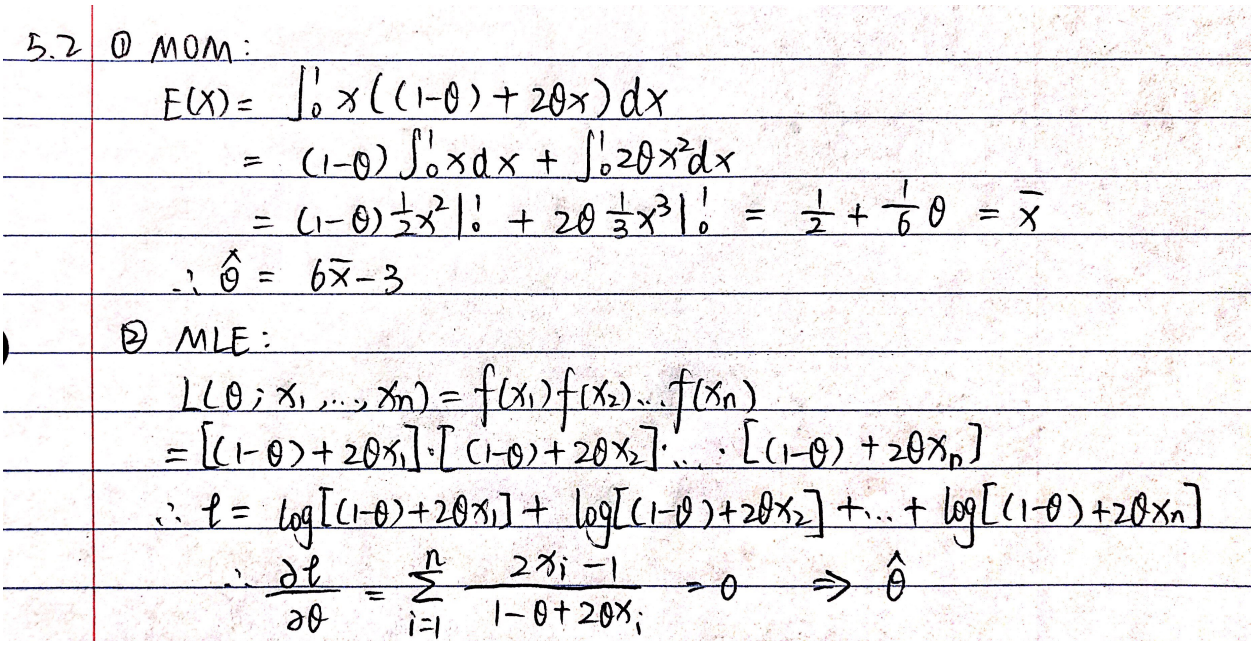
$$L(\lambda; x_1, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n) = \lambda^n e^{-\lambda \sum_{i=0}^{n} x_i}$$

$$\therefore \ell(\lambda; x_1, \ldots, x_n) = n\log(\lambda) - \lambda \sum_{i=0}^{n} x_i$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=0}^{n} x_i = 0$$

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

5.2 A new distribution



5.2 ① MOM:

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

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

$$= (1-\theta)\tfrac{1}{2}x^2\Big|_0^1 + 2\theta\tfrac{1}{3}x^3\Big|_0^1 = \tfrac{1}{2} + \tfrac{1}{6}\theta = \bar{x}$$

$$\therefore \hat{\theta} = 6\bar{x} - 3$$

② MLE:

$$L(\theta; x_1, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$$

$$= [(1-\theta) + 2\theta x_1]\cdot[(1-\theta) + 2\theta x_2]\cdot\ldots\cdot[(1-\theta) + 2\theta x_n]$$

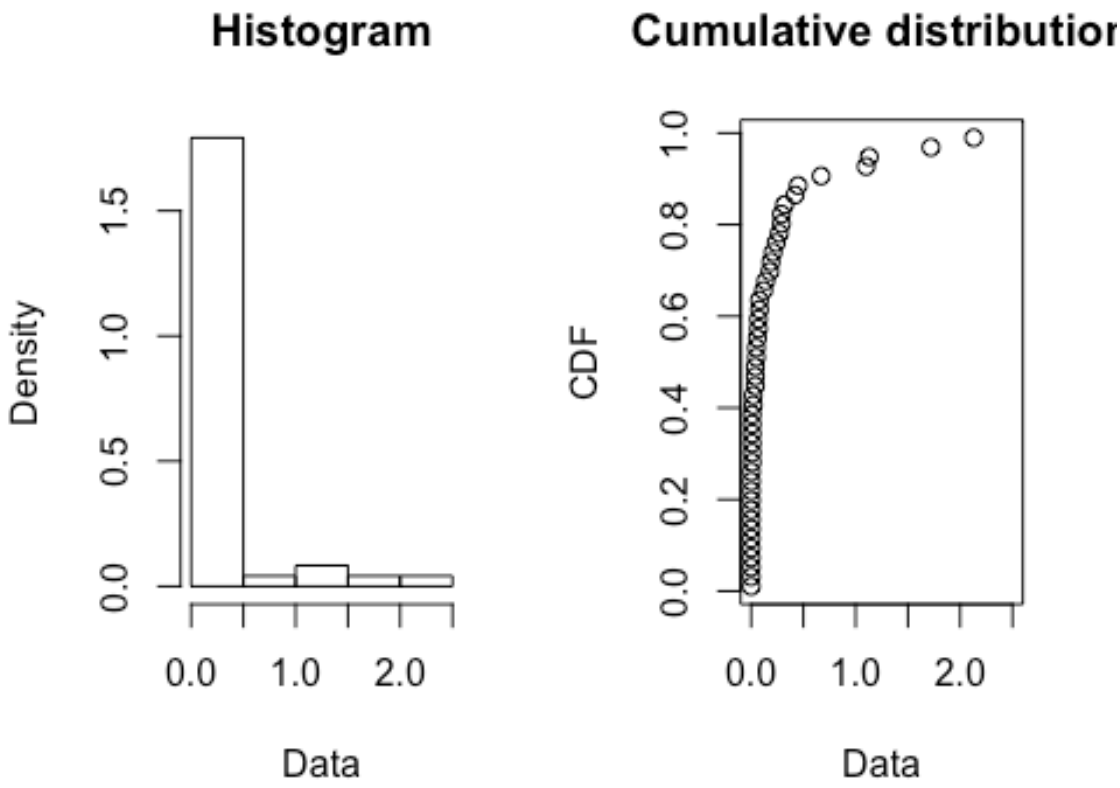
$$\therefore \ell = \log[(1-\theta) + 2\theta x_1] + \log[(1-\theta) + 2\theta x_2] + \ldots + \log[(1-\theta) + 2\theta x_n]$$

$$\therefore \frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{2x_i - 1}{1 - \theta + 2\theta x_i} = 0 \Rightarrow \hat{\theta}$$

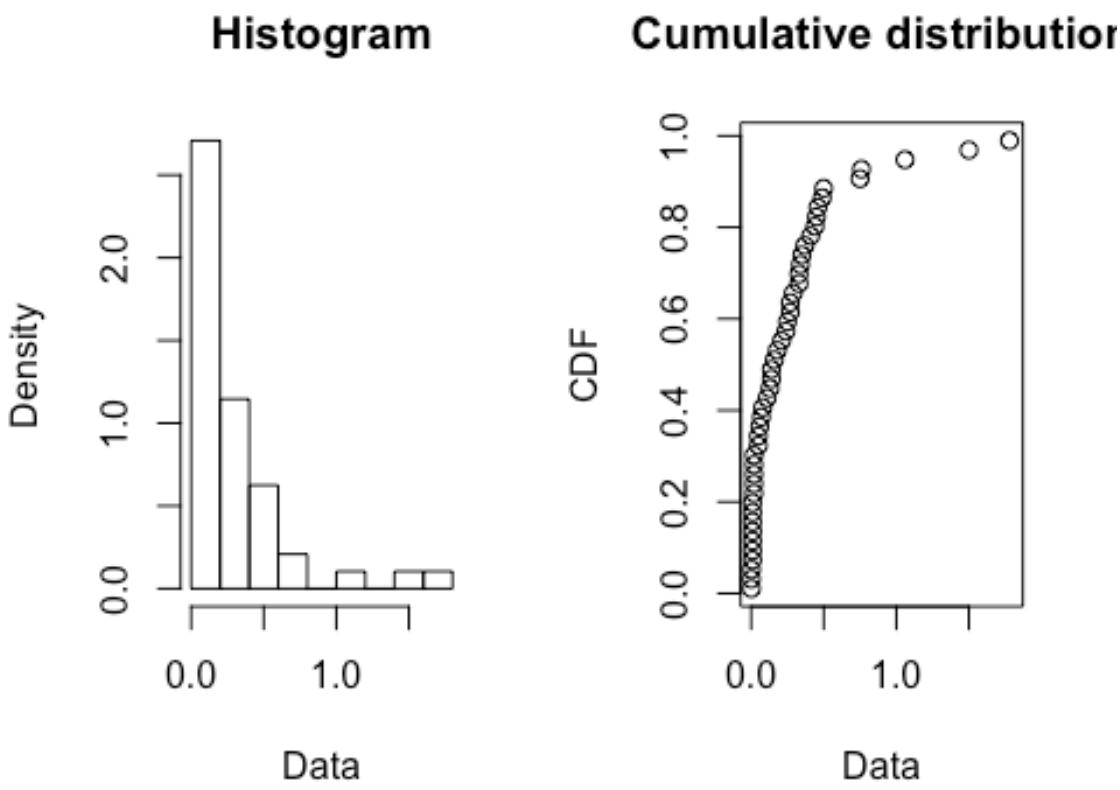
5.3 Rain in Southern Illinois

```r
# read the data
data60 <- read.table("ill-60.txt", quote="\"", comment.char="")
data60 <-as.numeric(as.array(data60 [,1]))
data61 <- read.table("ill-61.txt", quote="\"", comment.char="")
data61<-as.numeric(as.array(data61[,1]))
data62 <- read.table("ill-62.txt", quote="\"", comment.char="")
data62<-as.numeric(as.array(data62[,1]))
data63<- read.table("ill-63.txt", quote="\"", comment.char="")
data63<-as.numeric(as.array(data63[,1]))
data64 <- read.table("ill-64.txt", quote="\"", comment.char="")
data64<-as.numeric(as.array(data64[,1]))
```

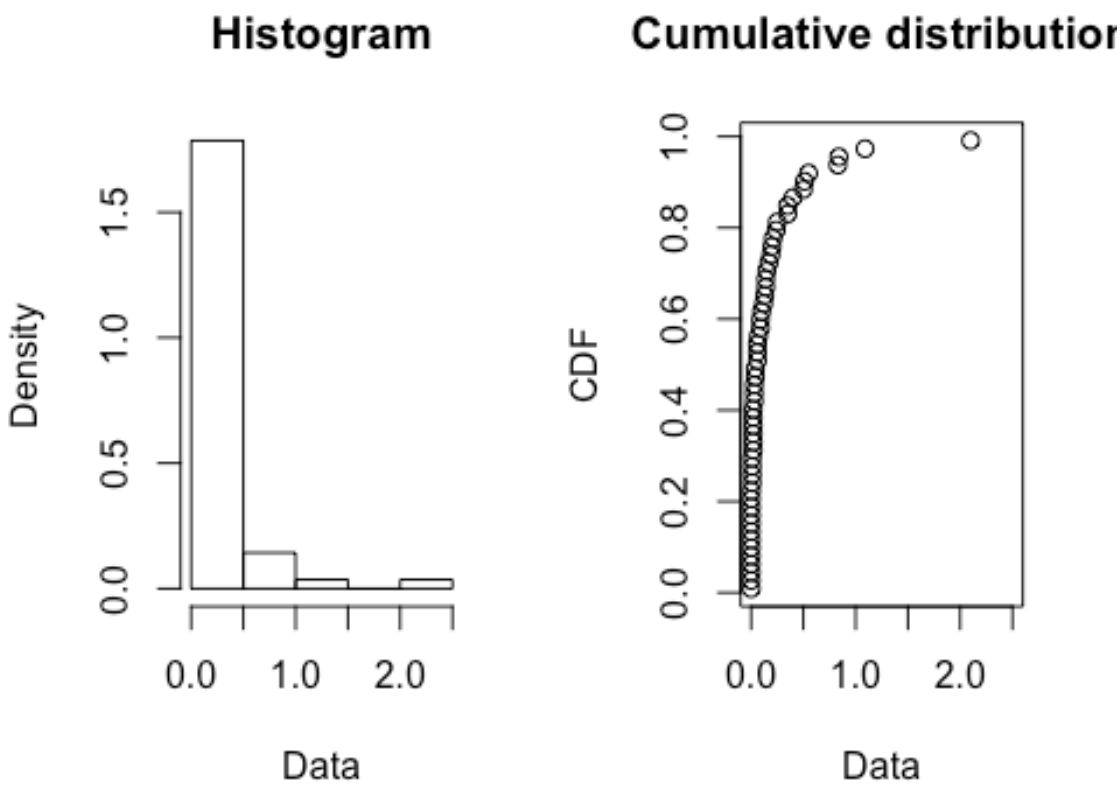
```r
# explore the distribution of the rainfall data
plotdist(data60)
```



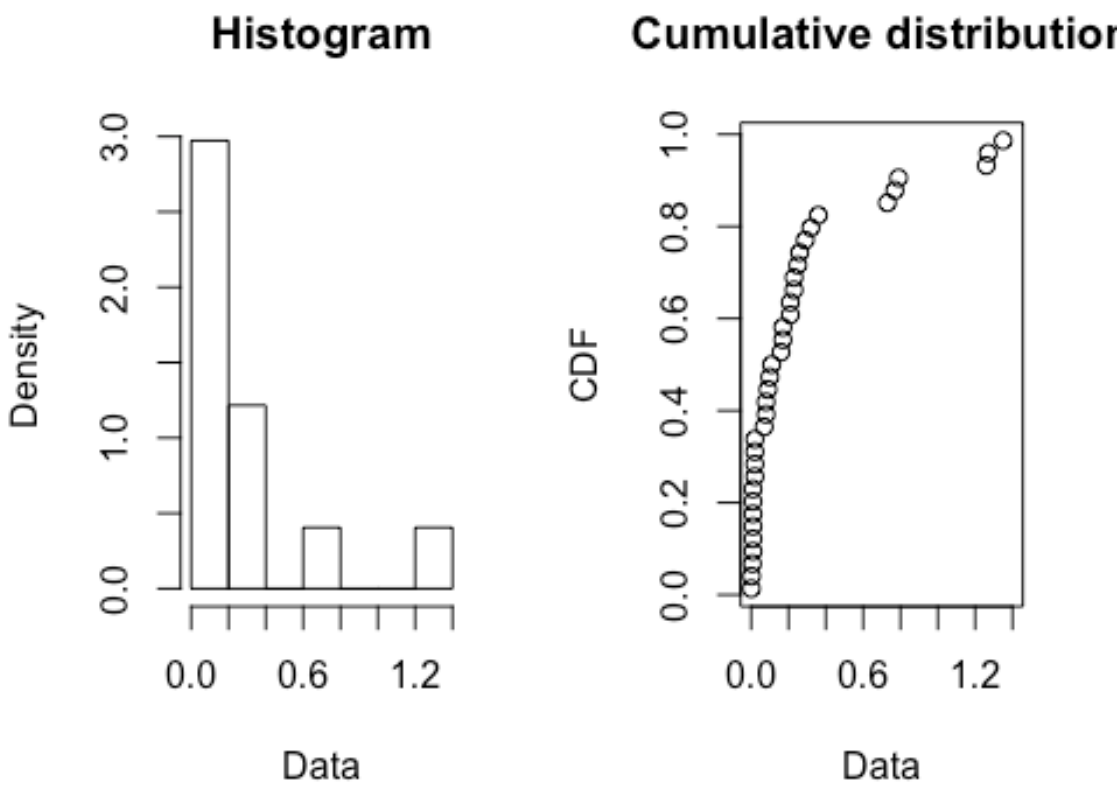

```r
plotdist(data61)
```



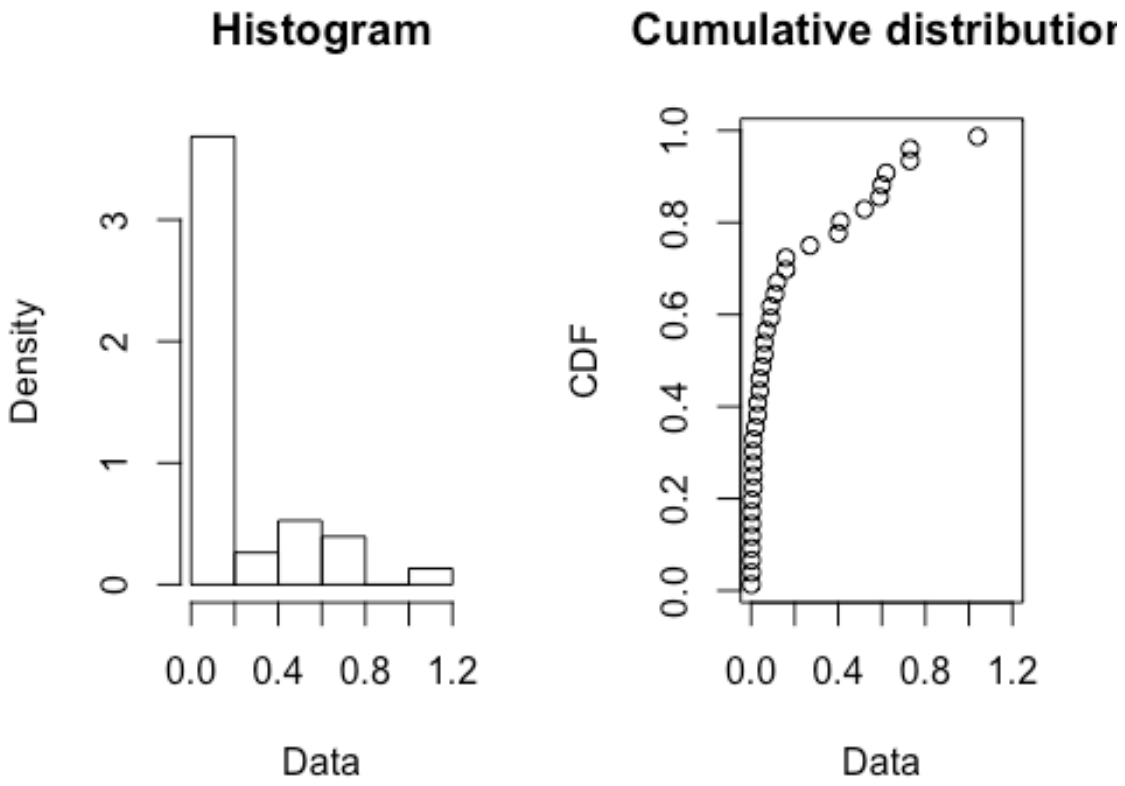

```
plotdist(data62)
```




```
plotdist(data63)
```




```
plotdist(data64)
```

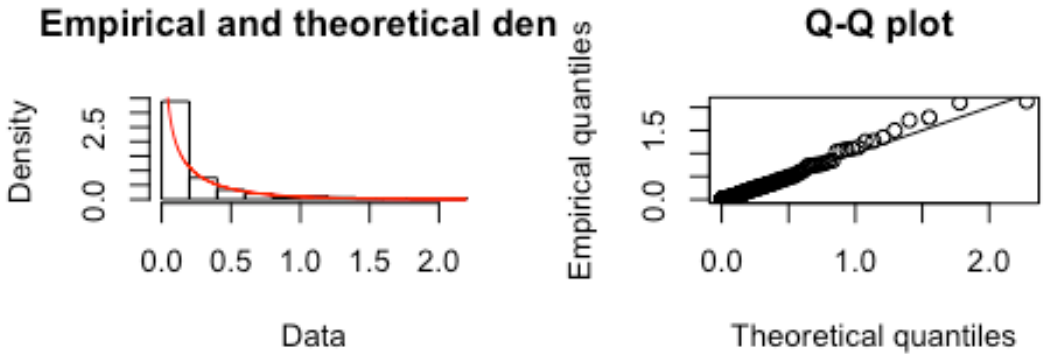


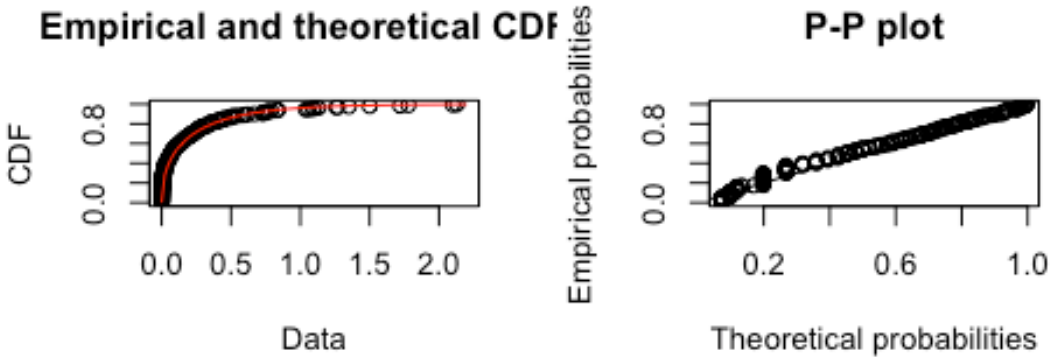

```
SumOfRain<-
as.data.frame(t(c(sum(data60),sum(data61),sum(data62),sum(data63),sum(data64)
)))
colnames(SumOfRain)[1:5]<-c("Total Rainfall in 1960","Total Rainfall in
1961","Total Rainfall in 1962","Total Rainfall in 1963","Total Rainfall in
1964")
kable(SumOfRain)
```

Total Rainfall in 1960	Total Rainfall in 1961	Total Rainfall in 1962	Total Rainfall in 1963	Total Rainfall in 1964
10.574	13.197	10.346	9.71	7.11

According to the distribution plot, five years are similar. 1961 is more wetter than others since it has the highest total rainfall.

```
#Test whether the gamma distribution was a good fit for their data.
alldata<-c(data60,data61,data62,data63,data64)
fgamma <- fitdist(alldata, "gamma")
plot(fgamma)
```





According to Q-Q plot and P-P plot, the gamma distribution was a good fit for their data. I totally agree with Changnon and Hu.

```
# calculate MOM and MLE
mom <- fitdist(alldata, "gamma",method = "mme")
boot_mom <- bootdist(mom)
summary(boot_mom)

## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.3967758 0.2680536 0.5326016
## rate  1.7743222 1.1481850 2.5228737

mle <- fitdist(alldata, "gamma",method = "mle")
boot_mle <- bootdist(mle)
summary(boot_mle)

## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4408721 0.3820586 0.5226438
## rate  1.9688032 1.5226623 2.5466529
```

For method of moment the 95% confidence interval of shape from bootstrap sample is (0.27,0.53), the rate is (1.17,2.58). For MLE, the 95% confidence interval of shape from

bootstrap sample is (0.38,0.51),the rate is (1.56,2.56). 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.

6. Analysis of decision theory article

Refer to:

Charles F. Manski (2019) Treatment Choice with Trial data: Statistical Decision Theory Should Supplant Hypothesis Testing, The American Statistician, 73:sup1, 296-304. Derive equations (10a), (10b), (10c) in Section 3.2.2. Use R to reproduce the calculations in Table 1 which is explained in 3.2.3.

Describe what you have done and what it means in the context the the treatment decision used as an illustration in the Manski article.

For any $\delta \in [0,1]$:

$$U(\delta,P) = R[y(A)](1-\delta) + E[y(B)]\delta \\ =\alpha(1-\delta)+\beta\delta = \alpha+(\beta-\alpha)\delta$$

where $\alpha \equiv E[y(A)]$ and $\beta \equiv E[y(B)]$.

$$U(\delta, P, \psi) = \alpha + (\beta - \alpha)\delta(\psi)$$

$$W(\delta, P_s, Q_s) = \alpha_s + (\beta_s - \alpha_s)E_s[\delta(\psi)]$$

where $E_s[\delta(\psi)] \equiv \int_{\psi} \delta(\psi) dQ_s(\psi)$.

Suppose that the success probability $\alpha \equiv P[y(A) = 1]$ and not success probability $\beta \equiv P[y(B) = 1]$, the expected wellfare of rule δ is:

$$W(\delta, P, N) = \alpha + (\beta - \alpha)E[\delta(n)].$$

n is distributed binomial $B[\beta, N]$, so

$$E[\delta(n)] = \sum_{i=0}^{N} \delta(i) f(n = i; \beta, N),$$

where $f(n = i; \beta, N) \equiv N![i!(N-i)!]^{-1}\beta^i(1-\beta)^{N-i}$ is the probability of i successes and $\beta_s \equiv P_s[y(b) = 1]$.

Thus, δ is admissible if and only if

$$\delta(n) = 0 \; for \; n < n_0$$

$$\delta(n) = \lambda \; for \; n = n_0$$

$$\delta(n) = 1, \; for \; n > n_0.$$

for some $0 \leq n_0 \leq N$ and $0 \leq \lambda \leq 1$.

Let $(\beta_s, s \in S) = (0,1)$ and let the prior be Beta with parameters (c, d).

Then the posterior mean for β is $(c + n)/(c + d + N)$.

Thus, we can drive the resulting Bayes rule is:

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

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

$$\delta(n) = 1 \; for \; (c + n)/(c + d + N) < \alpha.$$