MA 677 Final Project

Xiang XU

1 Statistics and the Law Argument

Difference between the rates of mortgage application refusals of white applicants and minority applicatns constitued evidence of discrimination.

Test (1) the data are sufficient evidence of discrimination to warrant corrective action and
(2) the data are not sufficient.

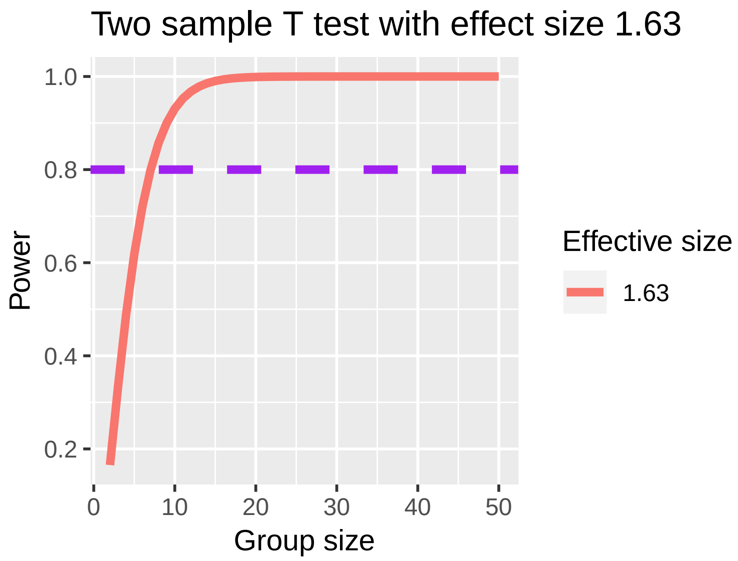
H_0: morgage refusal rate of white applicants are the same as that of minority applicant
H_1: morgage refusal rate of white applicants are lower than that of minority applicant

```
#Store the 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)#refusal rate for minority applicants
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)#refusal rate for white applicants
data1 <- data.frame(
                group = rep(c("Min", "White"), each = 20),
                percent = c(Min,  White)
                )

# Two sample T-test
# to test whether the refusal rate for minority applicants
# is greater than the refusal rate for white applicants
res <-  t.test(percent ~ group, data = data1,
        var.equal = TRUE, alternative = "greater")
res$p.value
```

```
## [1] 1.279668e-07
```

The p-value of this t test is about zero (1.279668e-07) – less than the significant level $\alpha = 0.05$. We can conclude that refusal rate for minority applicants is significantly different from refusal rate for white applicants with a p-value = 1.279668e-07.

```
# Power analysis to show the sufficiency
# Calculate the effect size
effect_size=abs(mean(Min)-mean(White))/sd(Min)

ptab1 <- cbind(NULL)
n <- seq(2, 50, by = 1)
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) +
  scale_y_continuous(breaks = seq(0, 1, by = 0.2)) +
  ggtitle("Two sample T test with effect size 1.63") +
  ylab("Power") +
  xlab("Group size")
```



```
# According to this power analysis plot, if we want to reach the general acceptable power 0.8, we need
```

2 Comparing Suppliers Revenue aside, which of the three schools produces the higher quality ornithopters, or are do they all produce about the same quality?

H_0 : They all produce about the same quality
H_1 : 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.

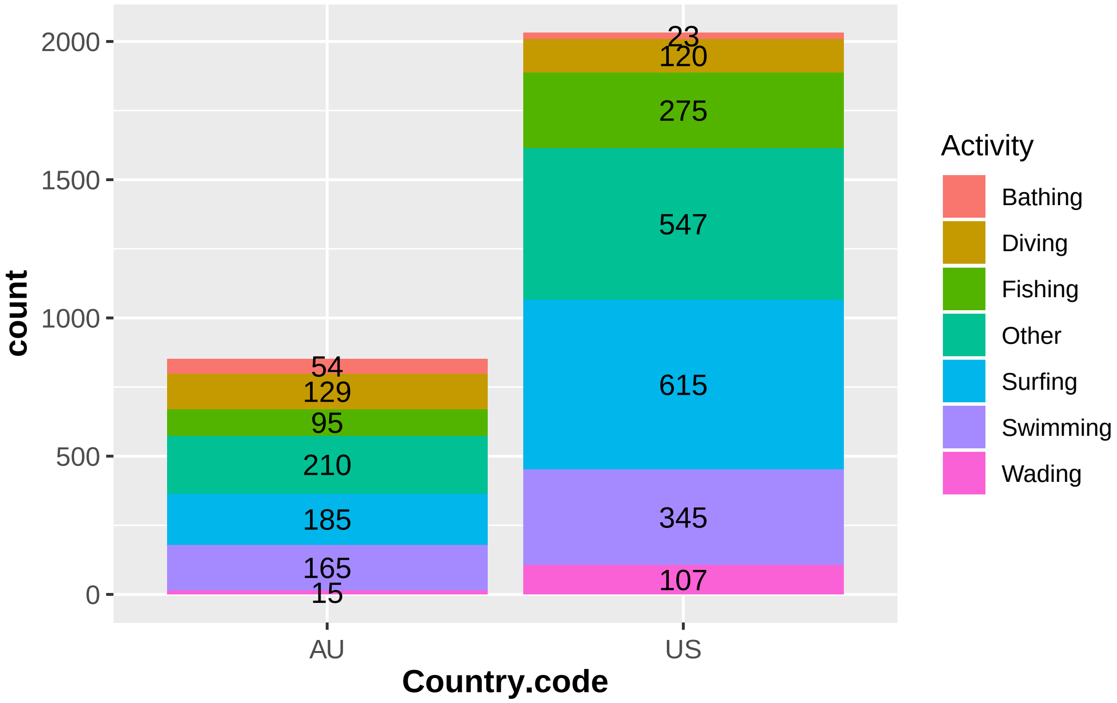
3 How deadly are sharks?

H_0: Sharks in Australia were, on average, are the same as the sharks in the United States
H_1: Sharks in Australia were, on average, are more vicious lot than the sharks in the United States

```
data3  <-  read.csv("sharkattack.csv")
# If we want to analyze the difference of viciousness between US sharks and AU sharks, we only care abo
data3  <-  data3%>%filter(Country.code=="US"|Country.code=="AU"&Type=="Unprovoked")
temp  <-  data3%>%group_by(Country.code,Activity)%>%summarise(count=n())%>% ungroup()%>%group_by(Countr
kable(temp)
```

Country.code	Activity	count	percent
AU	Bathing	54	0.0633060
AU	Diving	129	0.1512309
AU	Fishing	95	0.1113716
AU	Other	210	0.2461899
AU	Surfing	185	0.2168816
AU	Swimming	165	0.1934349
AU	Wading	15	0.0175850
US	Bathing	23	0.0113189
US	Diving	120	0.0590551
US	Fishing	275	0.1353346
US	Other	547	0.2691929
US	Surfing	615	0.3026575
US	Swimming	345	0.1697835
US	Wading	107	0.0526575

```
ggplot(data = temp, aes(Country.code, count, group = Activity)) +

 geom_col(aes(fill = Activity)) +

 geom_text(aes(label = count), position = position_stack(vjust = 0.5))+

  theme(axis.title.x = element_text(face="bold",  size=12),
        axis.title.y = element_text(face="bold",  size=12),
        plot.title = element_text(size=12, face="bold"),
        axis.text.x  = element_text(vjust=0.5, size=10),
        axis.text.y = element_text(vjust=0.5, size=10)) +

  theme(plot.title = element_text(hjust = 0.5))+ggtitle("")
```



According to the bar plot and data frame, US sharks made more attacks in total and attacks in surfing had higher percentage than AU sharks.

```
# Transfer dataframe into matric to do chi-square test
data33 <- matrix(c(23,120,275,547,615,347,107,54,129,95,210,186,165,12), nrow=2,
               dimnames = list(c("AU","US"),c("Bathing","Diving","Fishing","Other","Surfing","Swimming"
chisq.test(data33,correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data33
## X-squared = 378.78, df = 6, p-value < 2.2e-16
```

The p-value of this chi-square test is much smaller than the significant level α =0.05. Therefore, we reject the null hypothesis. Sharks in Australia were, on average, are more vicious than the sharks in the United States.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 . and then transformed t1 -t2 = h, which has euqal dectectability. This can solve the problem of falling into either side of the range.

5 Use the Method of Moments and MLE to find estimators as described in these three cases.

Case1 MLE of Exponential Distribution

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

$L(\lambda; x_1 \cdots x_n) = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda \sum x_i}$

$l(\lambda; x_1 \cdots x_n) = nlog(\lambda) - \lambda \sum x_i$

$\frac{\mathrm{d}l(\lambda; x_1 \cdots x_n)}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum x_i$

Thus, $\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x_n}}$

Case2

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

MOM

$E(x) = \int xf(x)dx = \int (x(1-\theta) + 2\theta x^2)dx$

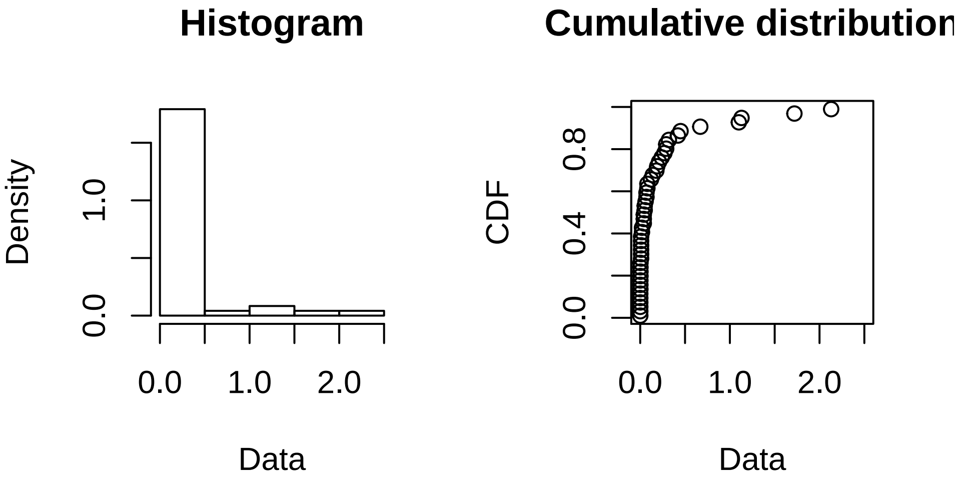
$= \int_0^1 (x - \theta x + 2\theta x^2)dx = \frac{1}{6}\theta + \frac{1}{2}$

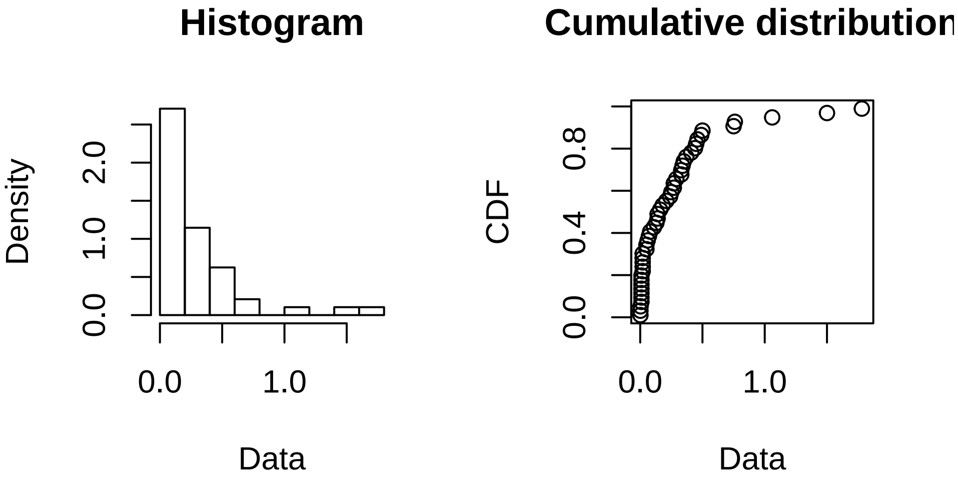
MLE

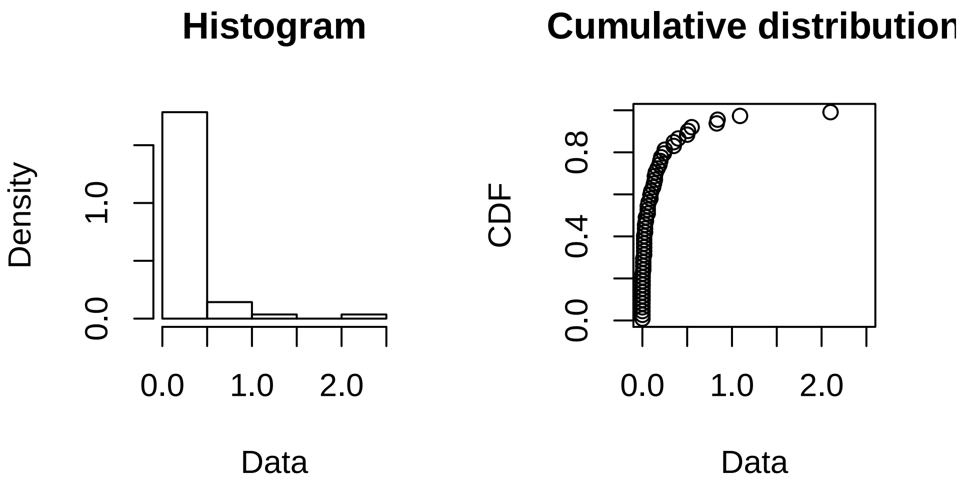
$L(\lambda; x_1 \cdots x_n) = \prod_{i=1}^{n} [1 - \theta + 2\theta x_i]$
$l(\lambda; x_1 \cdots x_n) = \sum log(1 - \theta + 2\theta x_i)$

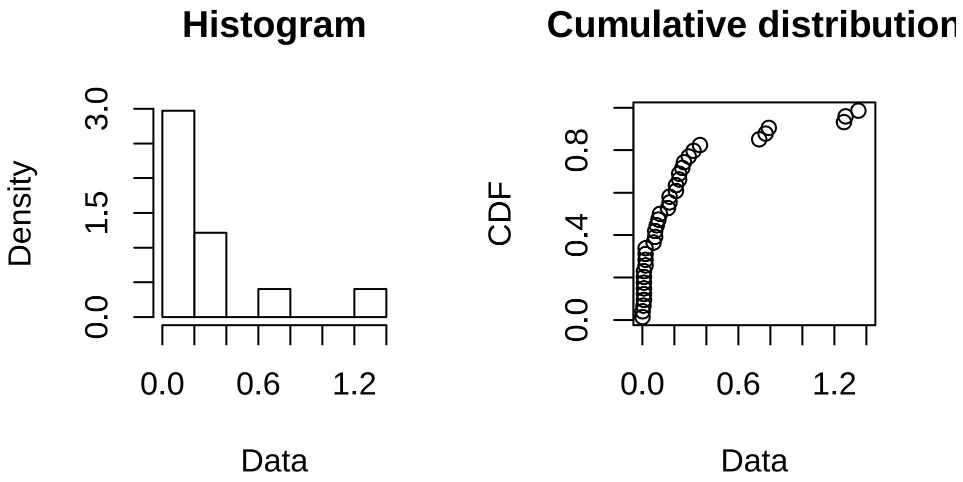
Case3

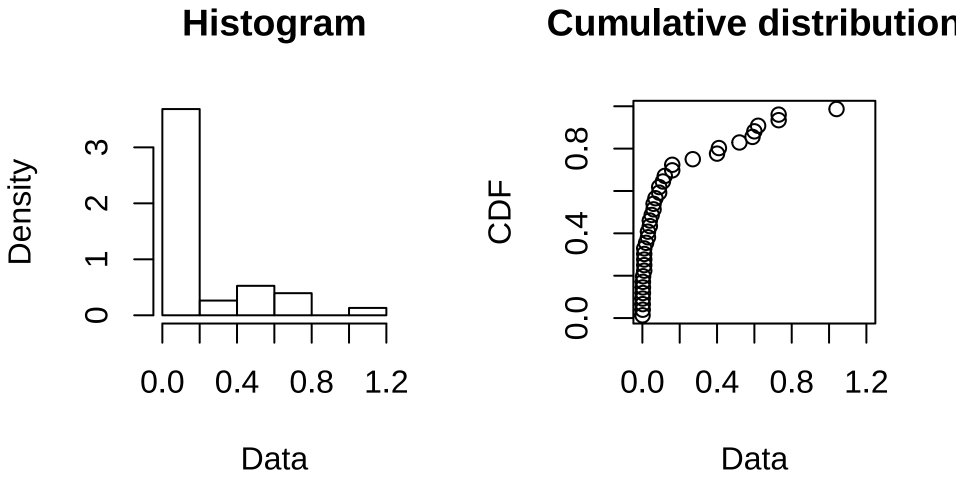
```r
# read the data

for(i in c(60:64)){

  path = paste0("ill-",i,".txt")

  if(file.exists(path)){
    # print(paste0(path, ' exists;'))
    data = read.table(path, quote="\"", comment.char="")

    # print(paste0("got data",path," ,with a dim ", dim(data)))
    data <- as.numeric(as.array(data[,1]))
    assign(paste0("df",i), data)
  }
  rm(data)
}


# explore the distribution of the rainfall data
plotdist(df60);plotdist(df61);plotdist(df62);plotdist(df63);plotdist(df64)
```


Histogram
Density
0.0
1.0
0.0
1.0
2.0
Data
Cumulative distribution
CDF
0.0
0.4
0.8
0.0
1.0
2.0
Data


Histogram
Density
0.0
1.0
2.0
0.0
1.0
Data
Cumulative distribution
CDF
0.0
0.4
0.8
0.0
1.0
Data


Histogram
Density
0.0
1.0
0.0
1.0
2.0
Data
Cumulative distribution
CDF
0.0
0.4
0.8
0.0
1.0
2.0
Data


Histogram
Density
0.0
1.5
3.0
0.0
0.6
1.2
Data
Cumulative distribution
CDF
0.0
0.4
0.8
0.0
0.6
1.2
Data

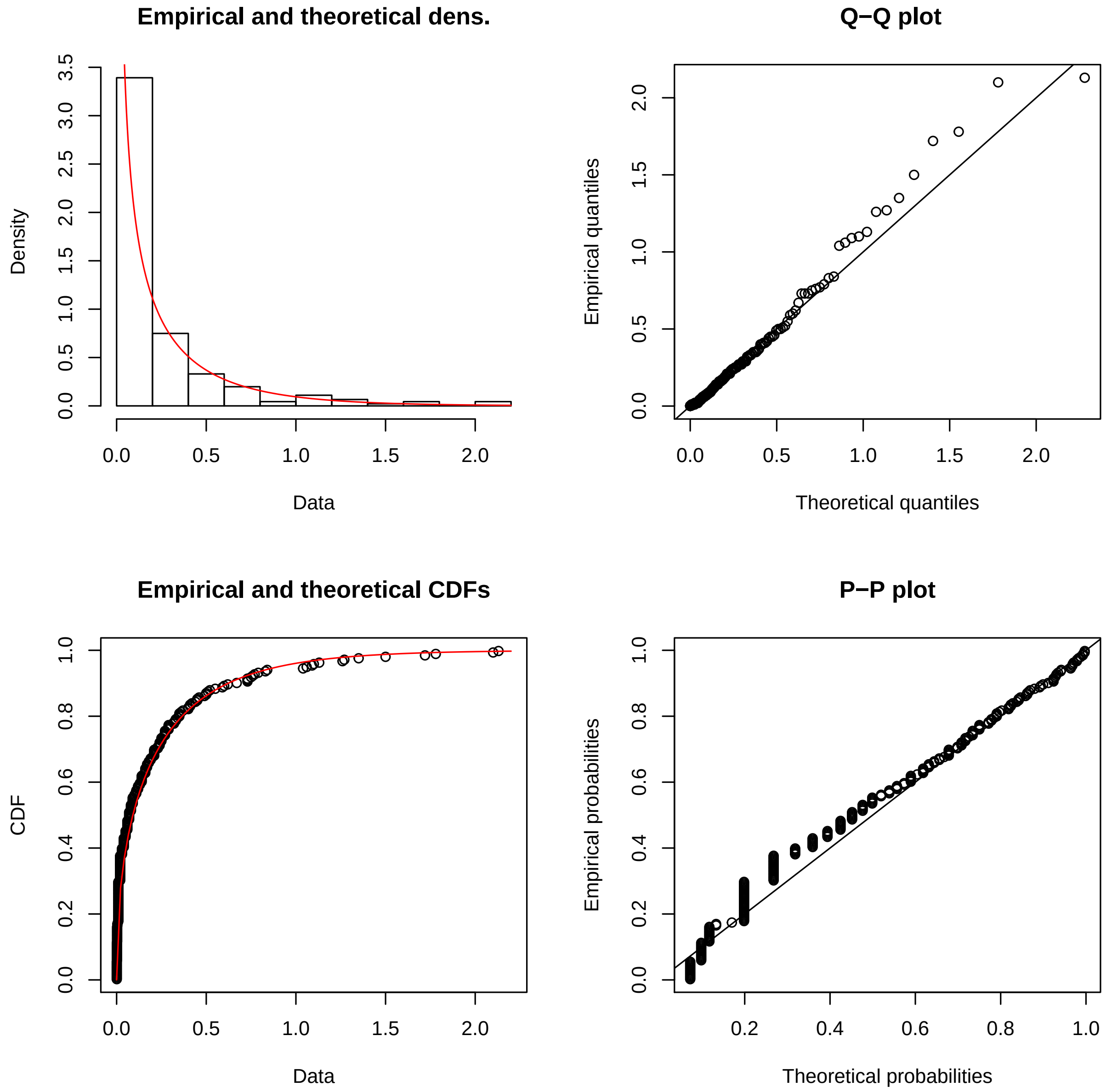


```
SumOfRain  <-  as.data.frame(t(c(sum(df60),sum(df61),sum(df62),sum(df63),sum(df64))))
colnames(SumOfRain)[1:5]  <-  c("Total Rainfall in 1960",
                                "Total Rainfall in 1961",
                                "Total Rainfall in 1962",
                                "Total 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 look similar. 1961 has the highest total rainfall.

```
#Test whether the gamma distribution was a good fit for their data.
df <- c(df60,df61,df62,df63,df64)
fgamma  <-  fitdist(df, "gamma")
plot(fgamma)
```



According to the Empirical the theoretical CDFs and P-P plot, the data points almost excellent fit the gamma distribution.Although in Q-Q plot, the points after 1.0 theoretical quantiles are a little bit discrete from the gamma distribution. Generally, the gamma distribution was a good fit for their data. I will agree with Changnon and Hu.

```
set.seed(2019)

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

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.3902039 0.2777425 0.5301561
## rate  1.7497653 1.1801093 2.5899000
```

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

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4447087 0.3808683 0.5170126
## rate  1.9913553 1.5616429 2.5430885
```

For MOM, 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 suggest MLE being the estimator because it has lower variance.

Analysis of decision theory article

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.