MA677_Final_Project

Chaoqun Yin

5/7/2019

1 Statistics and the Law

Question description: ACORN made a statstical arguement that the difference between the rates of mortgage application refusals of wihte applicants and minority applicatns constitued evidence of discrimination. Use ACORN's data and create the arguements that (1) the data are sufficient evidence of discrimination to warrant corrective action and (2) the data are not sufficient.

This question is equivalent to a test and a power analysis:

- H0: The refusal rate of white applicants equals the refusal rate of minority applicants.
- H1: The refusal rate of white applicants is higher than the refusal rate of minority applicants.

```
#prepare 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) #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) #white applicants
df_dis <- data.frame(group = rep(c("MIN", "WHITE"), each = length(MIN)),
                     percent = c(MIN,  WHITE))
#t-test
res_dis<-t.test(percent ~ group, data = df_dis,
         var.equal = TRUE, alternative = "greater")
res_dis$p.value #1.279668e-07

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

As we can see that the p-value of the t-test is pretty small (smaller than 0.05), so we should reject hypoyhesis 0. The conclusion is that the refusal rate of white applicants (WHITE) is higher than the refusal rate of minority applicants (MIN).

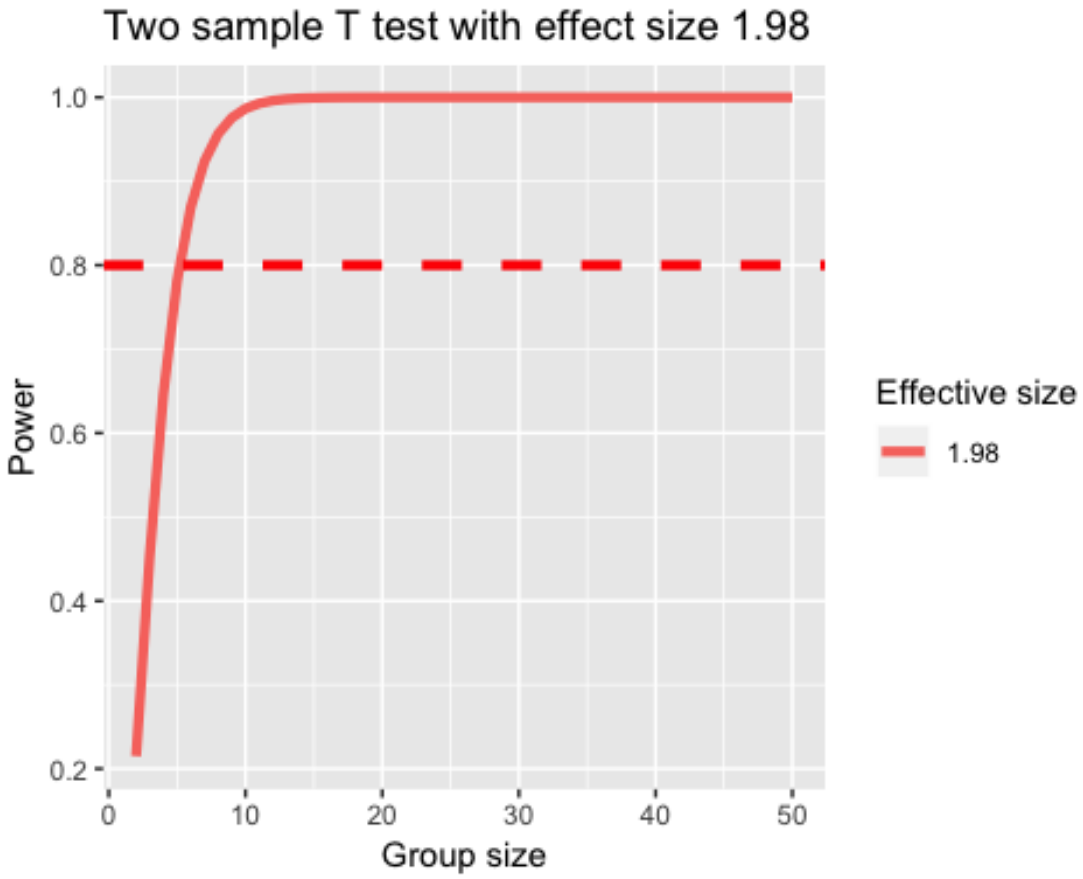
Then the following is a power analysis for confirming data's sufficiency.

```
#Calculate the effect size
sd_p <- sqrt((sd(WHITE) ^ 2 + sd(MIN) ^ 2) / 2)
eff_size <- abs((mean(WHITE) - mean(MIN)) / sd_p)

ptab1 <- cbind(NULL)
```

```r
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 = eff_size, alternative = "two.sided"
  )
  ptab1 <- rbind(ptab1, pwrt1$power)
}

ptab1<-as.data.frame(ptab1)
colnames(ptab1)[1]<-"num"
ggplot(ptab1) +
  geom_line(aes(x = n, y = num, colour = "orange"), size = 1.5) +
  scale_color_discrete(name = "Effective size",
                       labels =c(round(eff_size, 2))) +
  geom_hline(yintercept = 0.8, linetype = "dashed",
             color = "red", size = 1.5) +
  ylab("Power") +
  scale_y_continuous(breaks = seq(0, 1, by = 0.2)) +
  ggtitle("Two sample T test with effect size 1.98") +
  xlab("Group size")
```



Because of effect size of 1.98, an acceptable level of 0.8 requires more than 5 samples in each group, and we have 20 samples in our data in each group. Therefore, the data is sufficient for this case.

2 Comparing Suppliers

Question description: Acme Student Products sources ornithopters from high schools where students make orithopters as projects in a kinetics sculptor class. Not all of the ornithopers fly. Not all of them look good enough. Revenue aside, which of the three schools produces the higher quality ornithopters, or are do they all produce about the same quality?

This question is equivalent to this test below:

- H0: The three schools produce the same quality.
- H1: The three shcools produce different level of quality.

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

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

The chi-squared result shows that the p-value is extremely large (much larger than 0.05), which leads us to reject H0. The conclusion is that three schools produce the same quality.

3 How deadly are sharks?

Question description: Now that you have the data, please help me sort out how U.S. sharks compare with Australian sharks. Explain your analysis in terms that are simple but technically correct, make sure to include an analysis of statistical power.

For this case, it is equivalent to this test:

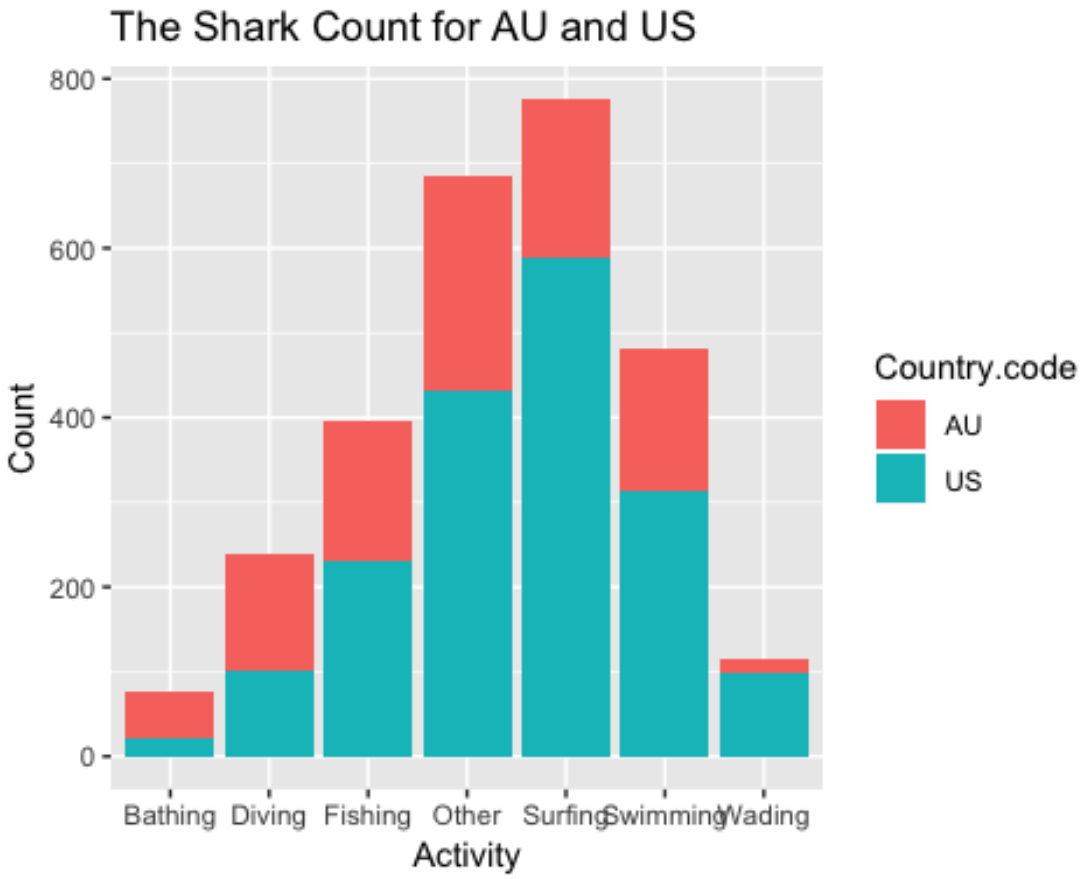
- H0: Sharks in Australia were, on average, are the same as the sharks in the United States.
- H1: Sharks in Australia were, on average, are more vicious than the sharks in the United States.

```
df_shark <- read.csv("material/sharkattack.csv")
#In this case we only need sharks in the US and the AU.
```

```
df_shark %>% filter(Country.code=="US" | Country.code=="AU") %>%
  filter(Type=="Provoked" | Type=="Unprovoked") -> shark_m
shark_m %>%
  group_by(Country.code, Activity) %>%
  summarise(count=n()) %>%
  ungroup() %>%
  group_by(Country.code) %>%
  mutate(percent=count/sum(count)) -> shark_f
kable(shark_f)
```

Country.code, Activity, count, percent
AU, Bathing, 55, 0.0560652
AU, Diving, 137, 0.1396534
AU, Fishing, 164, 0.1671764
AU, Other, 254, 0.2589195
AU, Surfing, 187, 0.1906218
AU, Swimming, 168, 0.1712538
AU, Wading, 16, 0.0163099
US, Bathing, 22, 0.0123043
US, Diving, 102, 0.0570470
US, Fishing, 231, 0.1291946
US, Other, 431, 0.2410515
US, Surfing, 590, 0.3299776
US, Swimming, 313, 0.1750559
US, Wading, 99, 0.0553691

```
#plot the stats
ggplot(shark_f) +
  geom_col(aes(x = Activity,y = count,fill = Country.code)) +
  ylab("Count") +
  ggtitle("The Shark Count for AU and US")
```



From this plot we can see that US sharks' attack incidents are more frequent than AU's.

```
shark_ff <- 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","Wading")))
chisq.test(shark_ff)

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

The result from chi-square test shows that p-value is smaller than 0.05, so H0 is rejected and the conclusion is that sharks in US are different from those in Australia. From empirical plot, we can see that sharks in US make attack more frequent.

4 Power analysis

Question description: In testing the parameter of a binomial distribution, 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. Explain the use of the arcsine transformation. How does it work? Why does it work?

As 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, in which way the hypothetical parameters of the distribution doesn not provide a scale of equal units of detectability because 0.25 and 0.05 are at the edge of the distribution.

After arcsine transformation transforming the proportional parameters to the scale of $-\pi/2$ to $\pi/2$. This can solve the problem of falling into the edge of the range.

5 Estimators

Use the Method of Moments and MLE to find estimators as described in these 3 cases.

(1) Exponential Distribution

$$X_1, \ldots, X_n$$

are independent draws from an exponential distribution, exp(λ). Find the MLE of λ.

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

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

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

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

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

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

(2) New Distribution

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

Find the method of moments estimator and MLE for θ.

Method of moments estimator:

$$\begin{aligned} E[X] &= \int_0^1 x((1-\theta) + 2\theta x)dx \\ &= (1-\theta)\int_0^1 xdx + \int_0^1 2\theta x^2 dx \\ &= (1-\theta)\frac{1}{2}x^2 \,|_0^1 + 2\theta\frac{1}{3}x^3 \,|_0^1 \\ &= \frac{1}{2} - \frac{1}{2}\theta + \frac{2}{3}\theta \\ &= \frac{1}{6}\theta + \frac{1}{2} \end{aligned}$$

MLE:

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

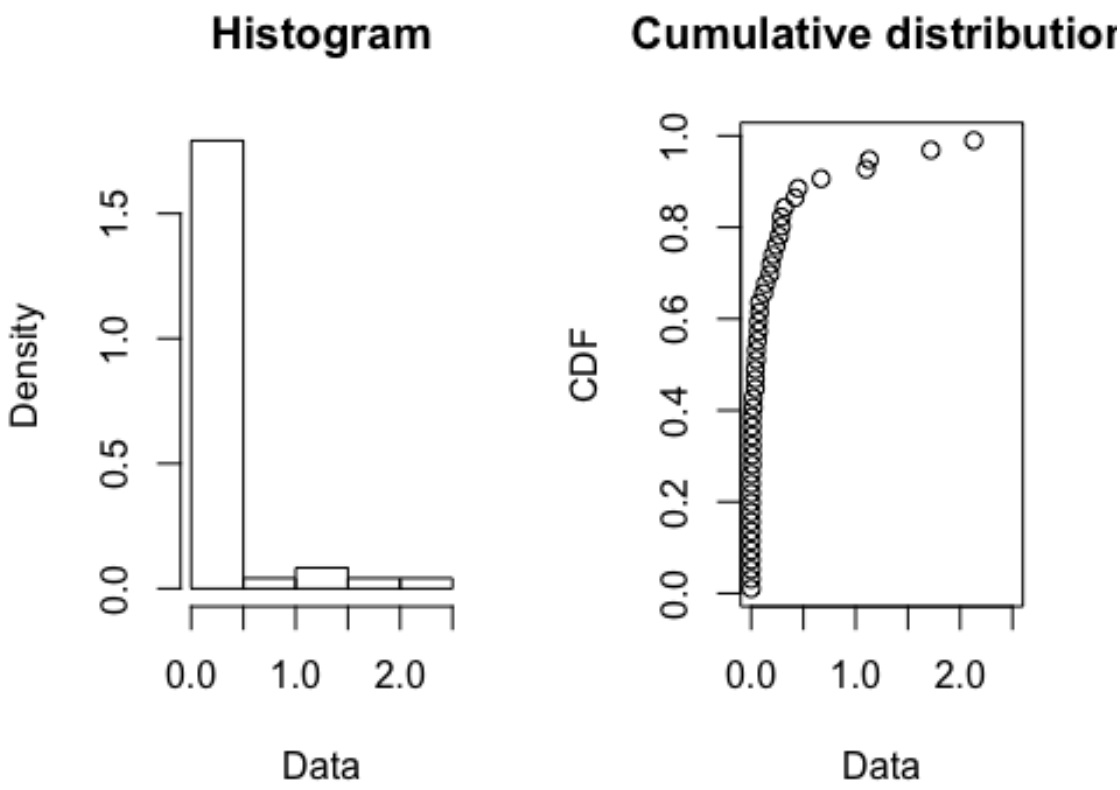
$$l(\theta; X_1, \ldots, X_n) = log[(1-\theta) + 2\theta X_1] + \ldots + log[(1-\theta) + 2\theta X_1]$$

Under this situation, it's not possible to find the maximum by taking the derivatives. We need to find θ to maximize in the other ways.

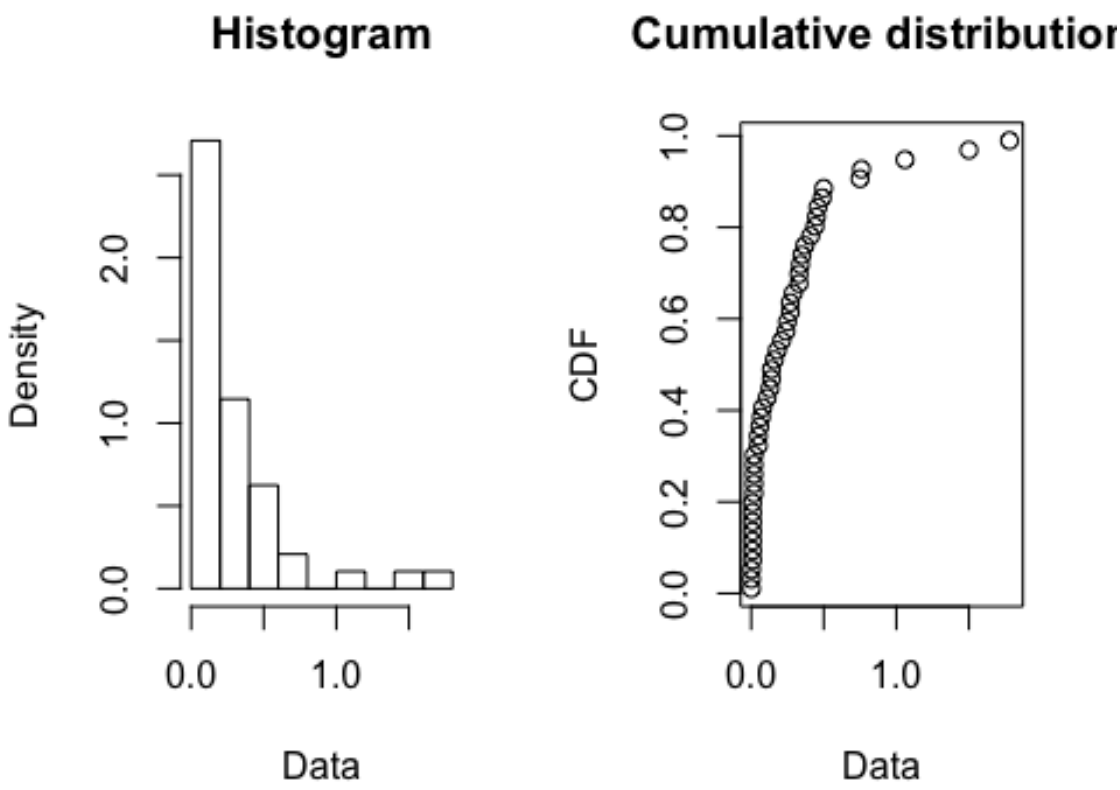
(3) Rain in Southern Illinois

In their article that Changnon and Huff concluded that the gamma distribution was a good fit for their data. What other distributions might they have considered? Do you agree with Changnon and Huff? Why? Why not? Using the gamma distribution as your model, produce estimates of the parameters using both the method of moments and maximum likelihood. Use the bootstrap to estimate the variance of the estimates. Compare the estimates which estimates would you present? Why?

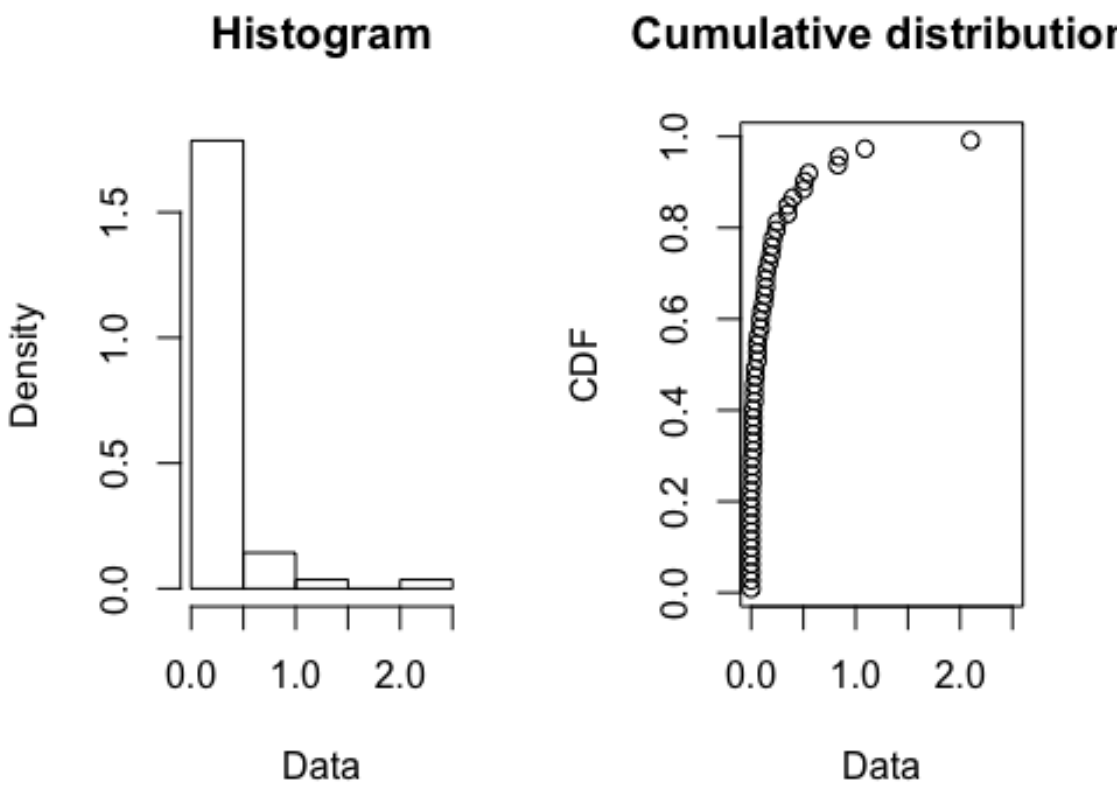
```
#read the data
data60 <- read.table("material/ill-60.txt", quote="\"", comment.char="")
data60 <-as.numeric(as.array(data60 [,1]))
data61 <- read.table("material/ill-61.txt", quote="\"", comment.char="")
data61<-as.numeric(as.array(data61[,1]))
data62 <- read.table("material/ill-62.txt", quote="\"",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



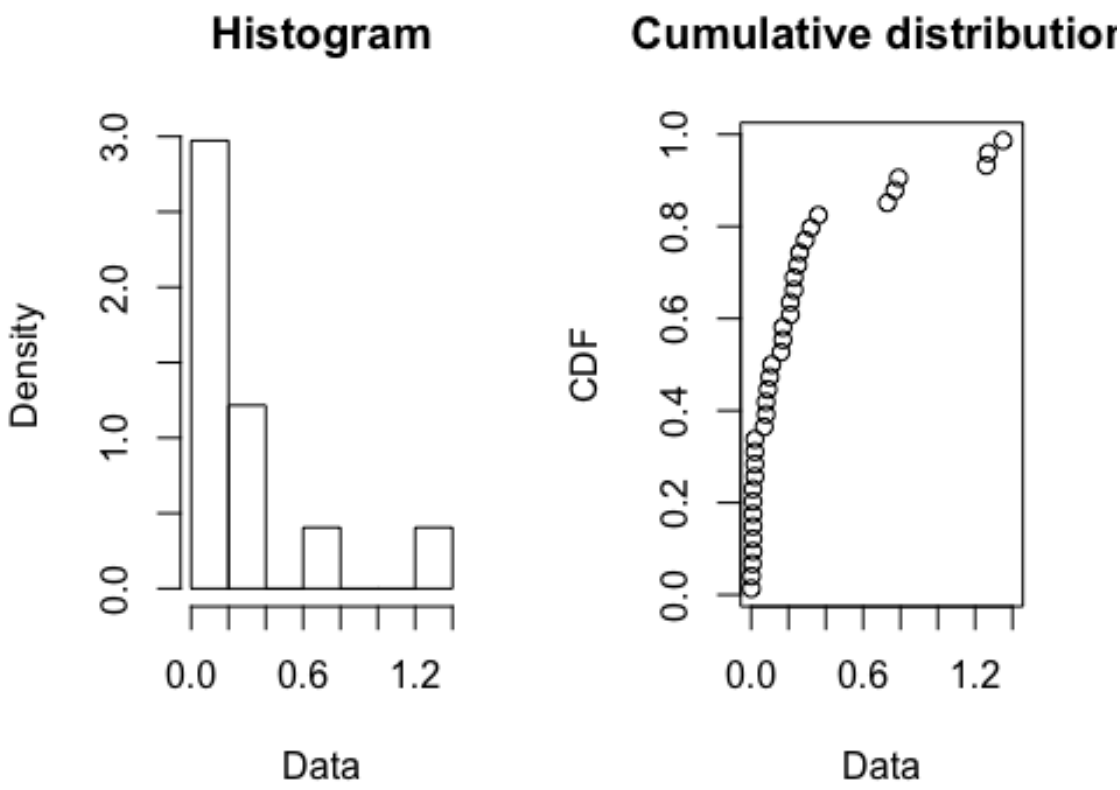

```
plotdist(data61)
```



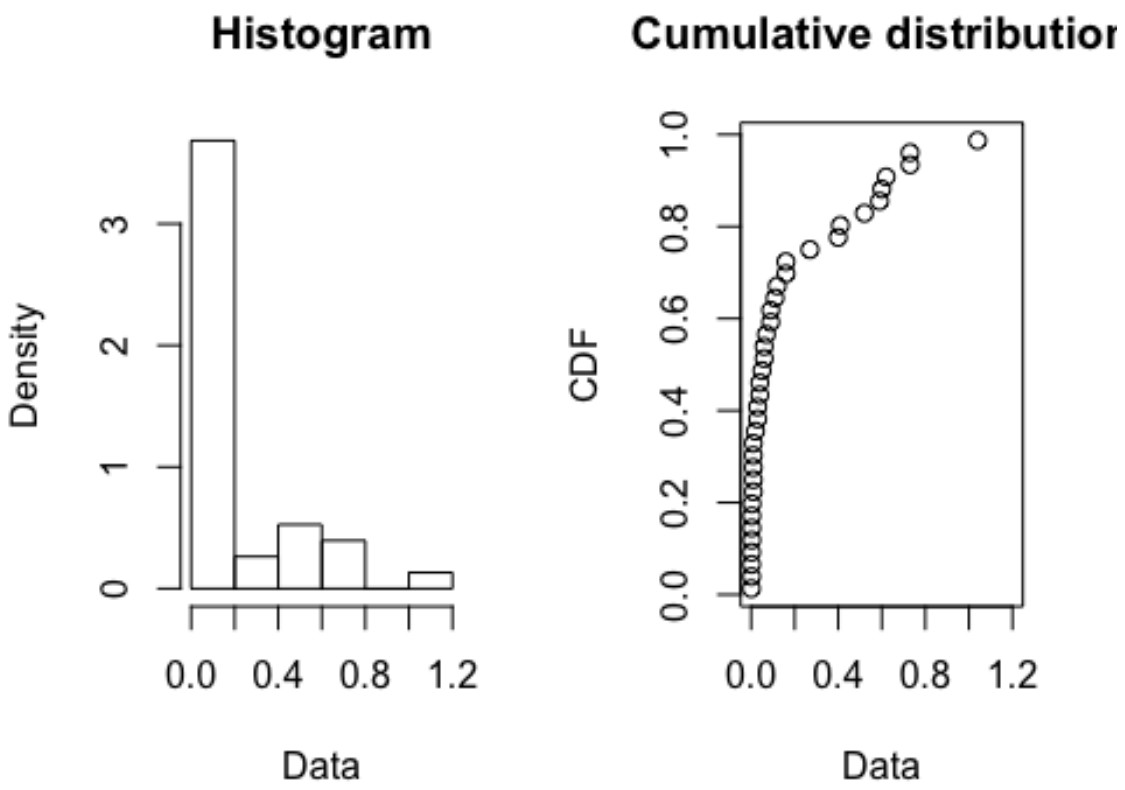

```
plotdist(data62)
```




```
plotdist(data63)
```




```
plotdist(data64)
```




```
sum(data60)
```

```
## [1] 10.574
```

```
sum(data61)
```

```
## [1] 13.197
```

```
sum(data62)
```

```
## [1] 10.346
```

```
sum(data63)
```

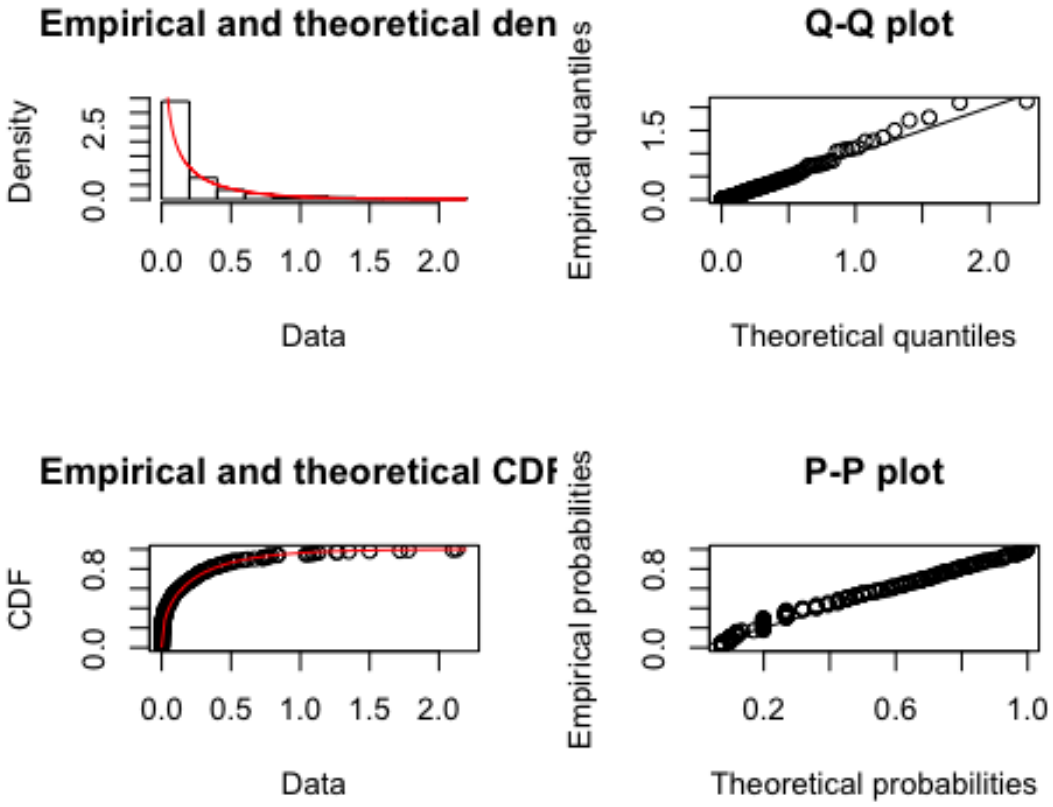
```
## [1] 9.71
```

```
sum(data64)
```

```
## [1] 7.11
```

The pecipitation amount in 1961 is the highest due to the storm.

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



From QQ-plot and PP-plot, it seems that gamma distribution is a good fit for the rain data. Therefore, Changnon and Huff is right.

```
#Compare the results from MLE and MOM
fmme <- fitdist(data_rain, "gamma",method = "mme")
bmme <- bootdist(fmme)
summary(bmme)

## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.3944301 0.2601882 0.5279496
## rate  1.7585961 1.1149540 2.5503923

fmle <- fitdist(data_rain, "gamma",method = "mle")
gmle <- bootdist(fmle)
summary(fmle)

## 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
```

The 95% confidence interval bootstrap MME result is (0.27,0.53) and the rate is (1.15,2.56), while the 95% confidence interval bootstrap MLE result is (0.38,0.51), and the rate is (1.57,2.59). Therefore, we will choose MLE as the estimator.