MA677 Final Project

Tianying Xu

2019/5/5

I. Statistics and Law: Is the data sufficient evidence of discrimination to warrant corrective action?

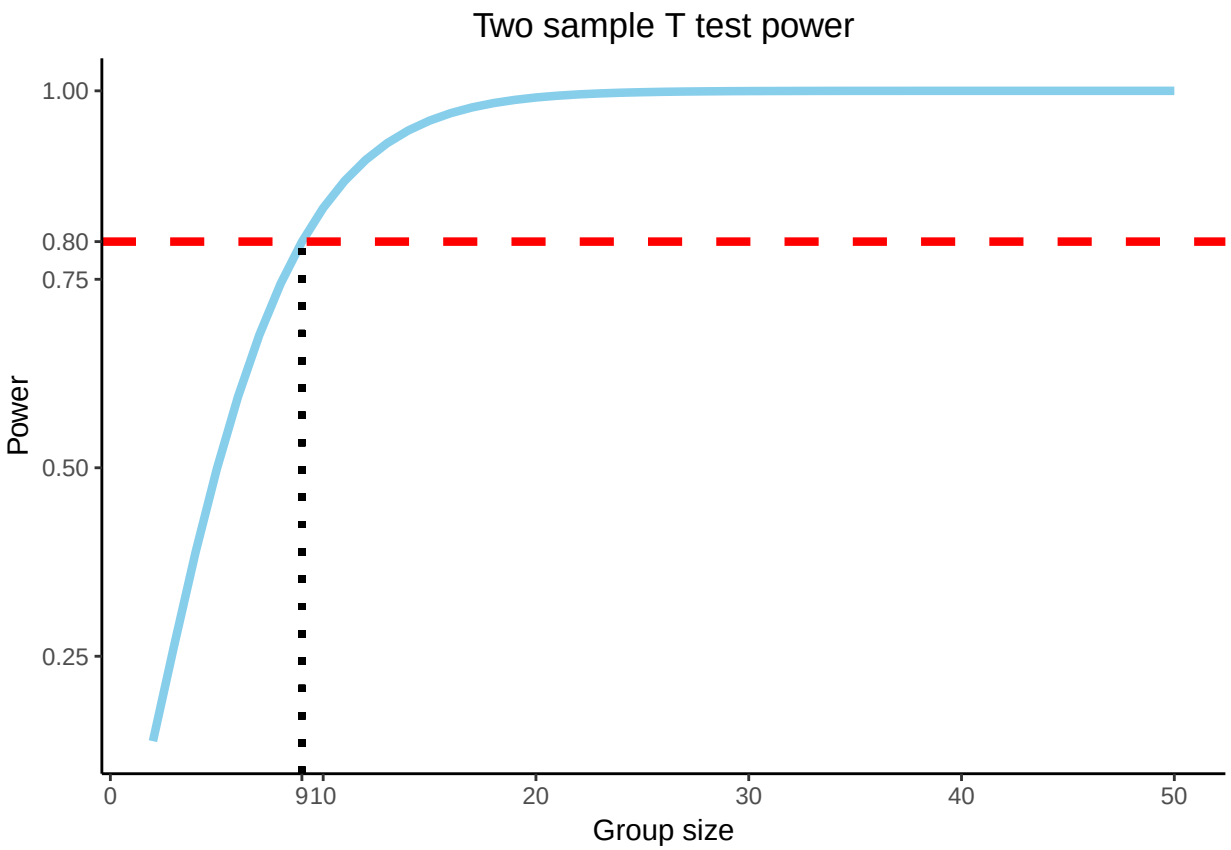
H0: morgage refusal rate for white applicants is the same as that of minority applicants H1: morgage refusal rate for white applicants is lower than that of minority applicants

```
D1 <- read.csv("acorn.csv",header=T)
y1=D1$MIN
y2=D1$WHITE
D1_ma <- t.test(y1,y2,paired = T)
D1_ma
```

```
##
##  Paired t-test
##
## data:  y1 and y2
## t = 11.46, df = 19, p-value = 5.619e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  17.37414 25.13886
## sample estimates:
## mean of the differences
##                  21.2565
```

```
# Power analysis to show the sufficiency
effect_size=abs(mean(y1)-mean(y2))/sd(c(y1,y2))
powers <- c()
ns <- 2:50
for (i in ns) {
  pwrt1 <- pwr.t2n.test(n1 = i, n2 = i,
                        sig.level = 0.05, power = NULL,
                        d = effect_size, alternative = "two.sided")
  powers <- rbind(powers, pwrt1$power)
}

D1_1 <- as.data.frame(ns)
D1_1$powers <- powers
ggplot(D1_1) +
  geom_line(aes(x = ns, y = powers), size = 1.5, colour = "skyblue") +
  scale_color_discrete(name = "Effective size", labels = c(round(effect_size,2))) +
  geom_hline(yintercept = 0.8, linetype = "dashed", color = "red", size = 1.5) +
  geom_segment(mapping=aes(x=9,y=-Inf,xend=9,yend=0.8),
               color="black",size=1.2,linetype="dotted")+
  ylab("Power") +
  theme_classic()+
  ggtitle("Two sample T test power") + xlab("Group size")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=c(0,9,10,20,30,40,50))+
  scale_y_continuous(breaks=c(0,0.25,0.5,0.75,0.8,1))
```



From paired t-test result, it is evident that there is statistically significant difference between these loan data of minority and white people. According to the power analysis plot, we need at least group size 9, which means 9 samples in each group. In the ACORN data, there is 20 samples in a group. Therefore, the data is sufficient evidence of discrimination to warrant corrective action.

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

H0:All three schools produce about the same quality H1:Three schools do not produce about the same quality

```
D2 <- matrix(c(12,23,89,8,12,62,21,30,119),ncol=3,nrow = 3,byrow=TRUE)
colnames(D2) <- c("Dead_Bird","Display_Art","Flying_Art")
rownames(D2) <- c("Area51","BDV","Giffen")
chisq.test(D2,correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  D2
## X-squared = 1.3006, df = 4, p-value = 0.8613
```

The Chi-Square test result shows that p-vlue is 0.8613, which is much larger than 0.05, therefore we can not reject the hypothesis. As a result, according to the chi-square test, all three schools produce about the same quality.

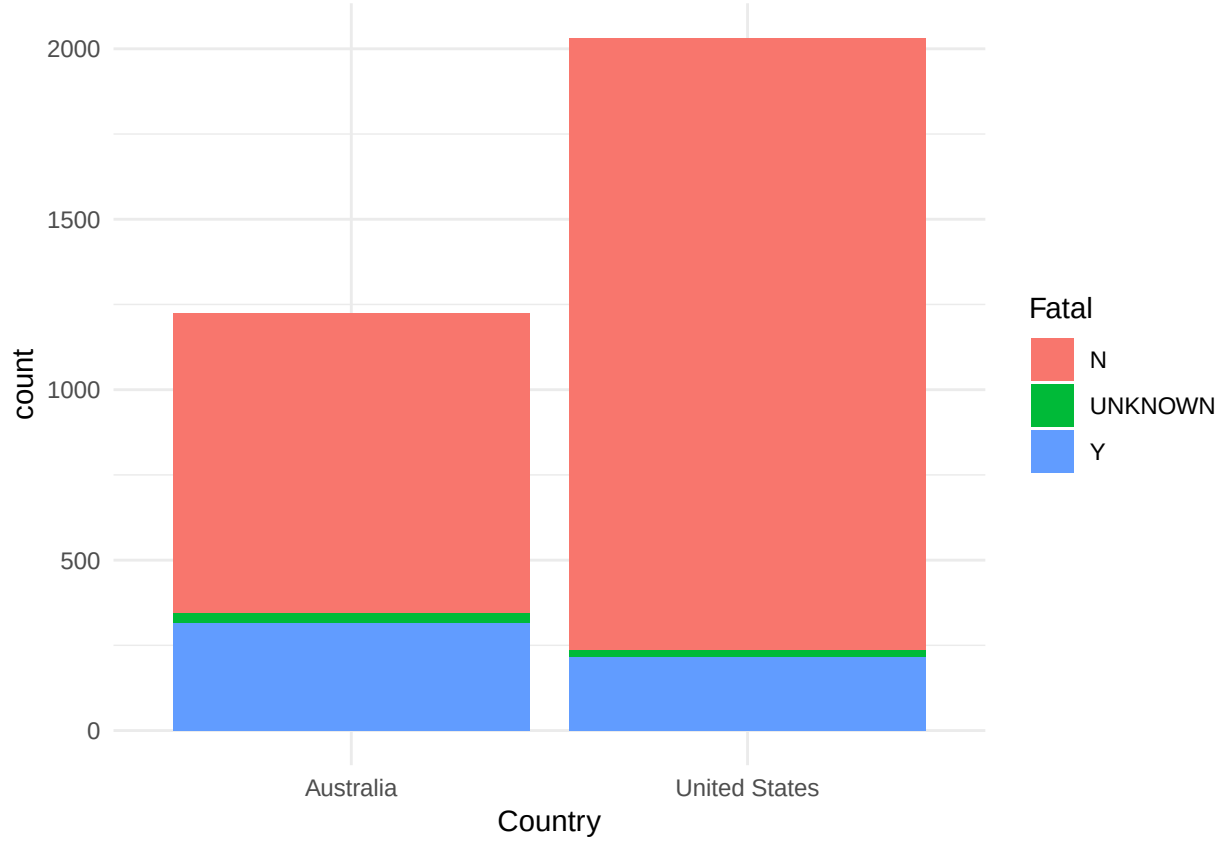
III. How deadly are sharks?

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

```
D3 <- read.csv("sharkattack.csv",header=T)
D3 <- D3 %>%
  filter(Country.code=="US" | Country.code=="AU")

ggplot(data=D3)+
  geom_bar(mapping=aes(x=Country, fill=Fatal, position = "stack"))+
  theme_minimal()
```

```
## Warning: Ignoring unknown aesthetics: position
```




```
D3_1 <- D3 %>%
  group_by(Country ,Country.code, Fatal) %>%
  summarise(n=n())
p1 <- D3_1$n[1:3]
p2 <- D3_1$n[4:6]

dd3 <- matrix(c(p1,p2),byrow = T, nrow = 2)
chisq.test(dd3,correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dd3
```

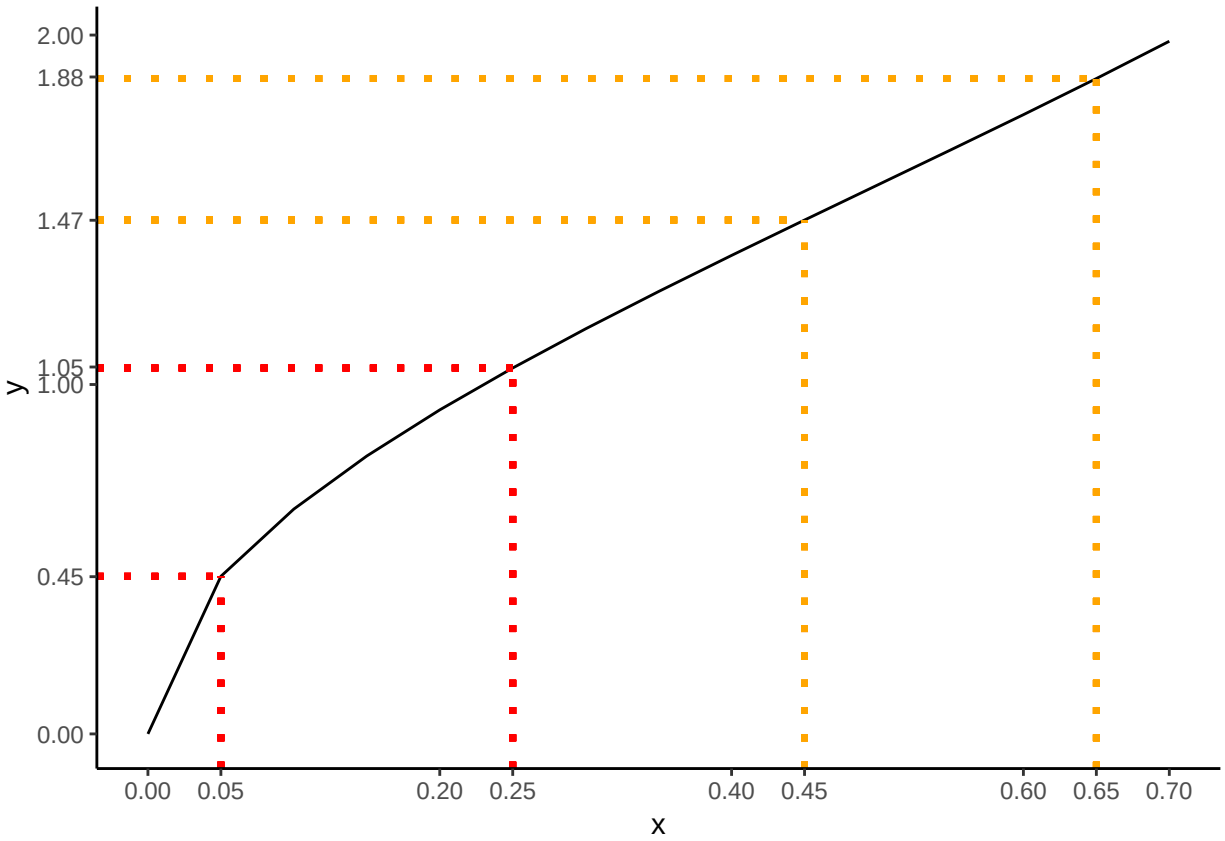
```
## X-squared = 142.13, df = 2, p-value < 2.2e-16
```

The bar plot indicate that sharks in Australia may be more vicious than sharks in the United States since although there are less incidents in Australia, the amount of fatal incident is more than in U.S. Also, from the chi-square test, the p-value is less than 0.05, which means that the hypothesis is rejected. Therefore, sharks in Australia were, on average, more vicious than sharks in the United States.

IV. Power Analysis

```
x <- seq(0,0.7,0.05)
y <- 2*asin(sqrt(x))
D=as.data.frame(matrix(c(x,y),byrow=F,ncol = 2))
names(D)=c("x","y")

ggplot(D)+
  geom_line(mapping=aes(x=x,y=y),color="black")+
  geom_segment(mapping=aes(x=0.05,y=-Inf,xend=0.05,yend=2*asin(sqrt(0.05))),
               color="red",size=1.2,alpha=0.8,linetype="dotted")+
  geom_segment(mapping=aes(x=-Inf,y=2*asin(sqrt(0.05)),xend=0.05,yend=2*asin(sqrt(0.05))),
               color="red",size=1.2,alpha=0.8,linetype="dotted")+
  geom_segment(mapping=aes(x=0.25,y=-Inf,xend=0.25,yend=2*asin(sqrt(0.25))),
               color="red",size=1.2,alpha=0.8,linetype="dotted")+
  geom_segment(mapping=aes(x=-Inf,y=2*asin(sqrt(0.25)),xend=0.25,yend=2*asin(sqrt(0.25))),
               color="red",size=1.2,alpha=0.8,linetype="dotted")+
  geom_segment(mapping=aes(x=0.45,y=-Inf,xend=0.45,yend=2*asin(sqrt(0.45))),
               color="orange",size=1.2,alpha=0.8,linetype="dotted")+
  geom_segment(mapping=aes(x=-Inf,y=2*asin(sqrt(0.45)),xend=0.45,yend=2*asin(sqrt(0.45))),
               color="orange",size=1.2,alpha=0.8,linetype="dotted")+
  geom_segment(mapping=aes(x=0.65,y=-Inf,xend=0.65,yend=2*asin(sqrt(0.65))),
               color="orange",size=1.2,alpha=0.8,linetype="dotted")+
  geom_segment(mapping=aes(x=-Inf,y=2*asin(sqrt(0.65)),xend=0.65,yend=2*asin(sqrt(0.65))),
               color="orange",size=1.2,alpha=0.8,linetype="dotted")+
  scale_x_continuous(breaks=c(0,0.05,0.2,0.25,0.4,0.45,0.6,0.65,0.7))+
  scale_y_continuous(breaks=c(0,round(2*asin(sqrt(0.05)),2),round(2*asin(sqrt(0.25)),2),round(2*asin(sq
  theme_classic()
```



According to Chapter 6 in the Jacob Cohen's Statistical Power Analysis for Behavioral sciences book, arcsine transformation on proportion is used in the pwr function. The formula is: $2 * arcsin(\sqrt{p})$.

This plot above indicate the function of arcsin transformation. In the plot, x is original proportion value and y is value after arcsin transformation. Although differences between (0.05,0.25) and (0.45,0.65) are both 0.2, differences after arcsin transformation become 0.6 and 0.41. Therefore, the power to detect the difference between hypothetical parameters .25 and .05 is .82, which is large than in the 0.65 and 0.45 case (0.48). This is all due to arcsin transformation, which makes the difference different.

V. Estimators

Case 1: Exponential

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

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

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

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

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

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

Case 2: New Distribution - Method of Moment

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

Case 3: New Distribution - MLE

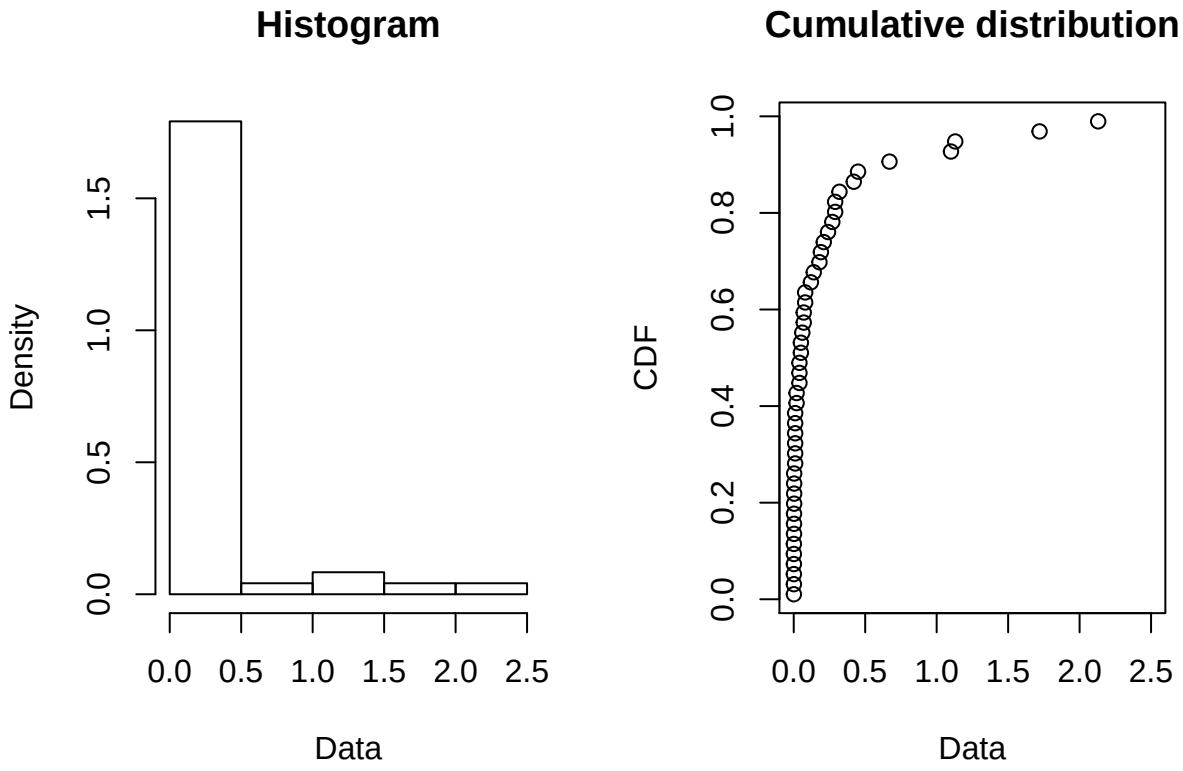
$$L(\theta; X_1, ..., X_n) = [(1-\theta) + 2\theta X_1]...[(1-\theta) + 2\theta X_n]$$

Rain in Southern Illinois

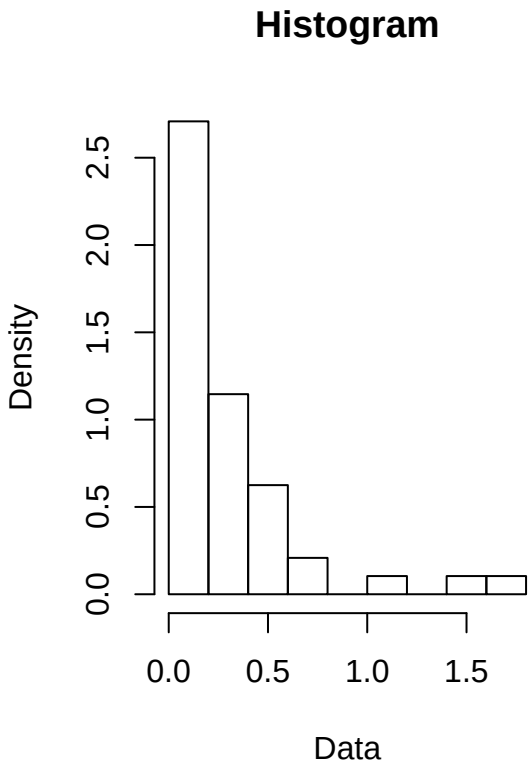
1. Are 5 years similar?

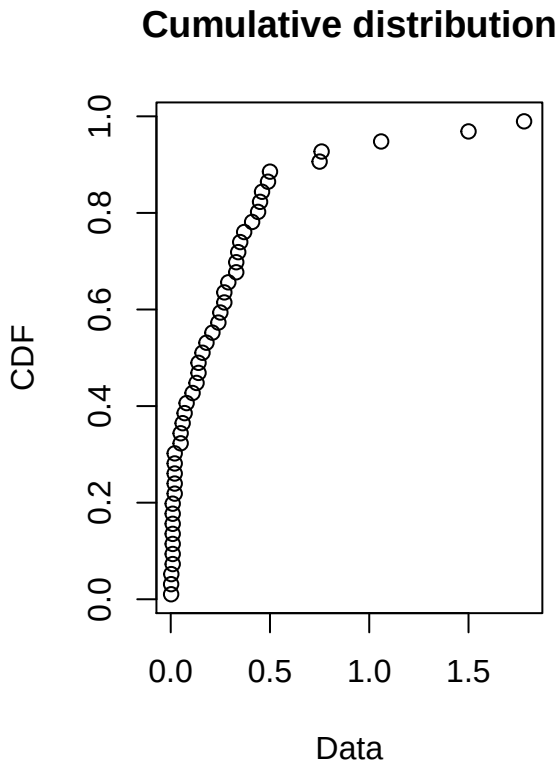
```
D6_0 <- read.table("ill-60.txt")
D6_1 <- read.table("ill-61.txt")
D6_2 <- read.table("ill-62.txt")
D6_3 <- read.table("ill-63.txt")
D6_4 <- read.table("ill-64.txt")
names(D6_0)=c("rain")
names(D6_1)=c("rain")
names(D6_2)=c("rain")
names(D6_3)=c("rain")
names(D6_4)=c("rain")
D6 <- rbind(D6_0,D6_1,D6_2,D6_3,D6_4)
D6$year <- c(rep("1960",48),rep("1961",48),rep("1962",56),rep("1963",37),rep("1964",38))
names(D6)=c("rain","year")

#Similar?Wetter?
plotdist(D6_0$rain)
```

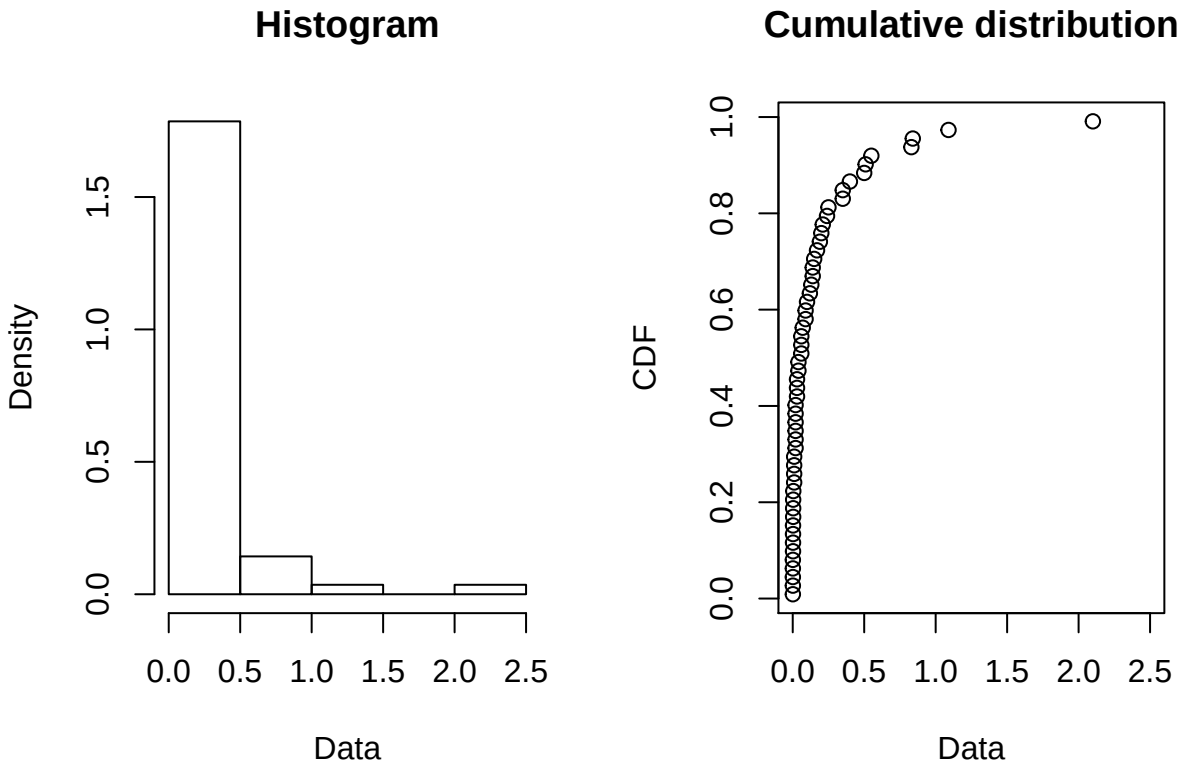



```
plotdist(D6_1$rain)
```

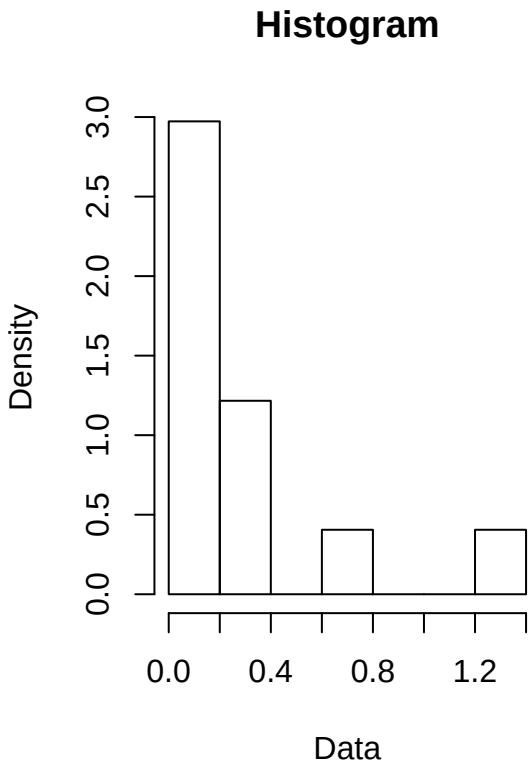


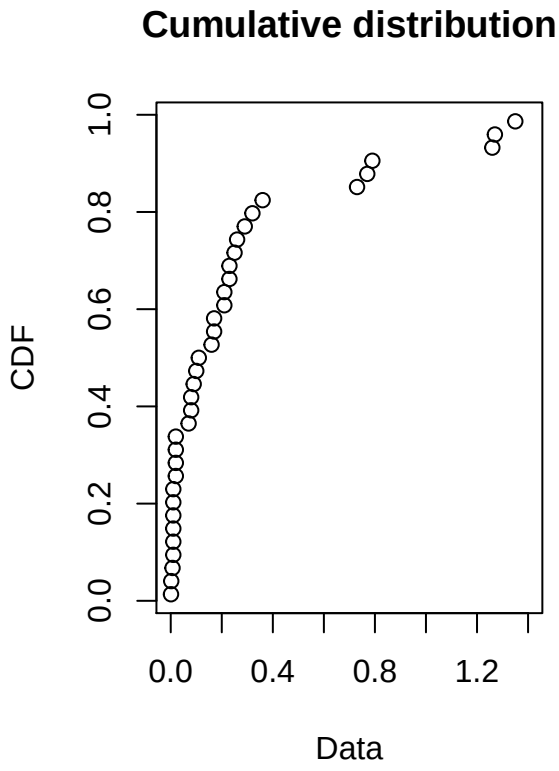



```
plotdist(D6_2$rain)
```

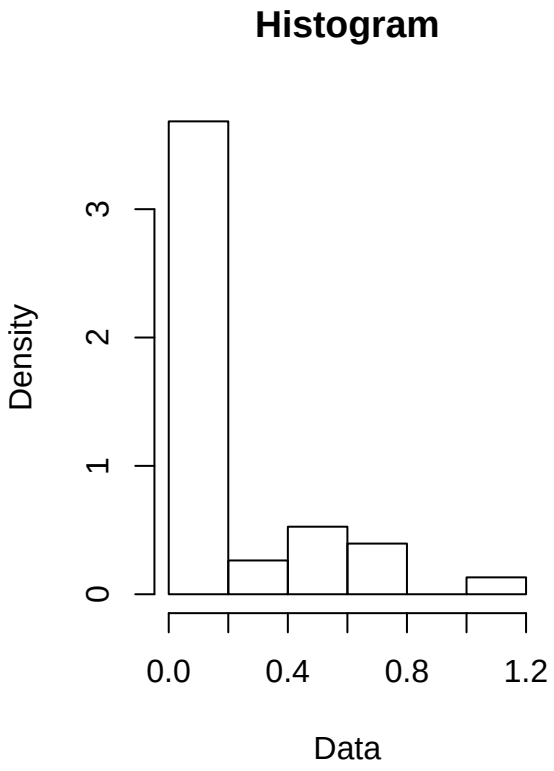


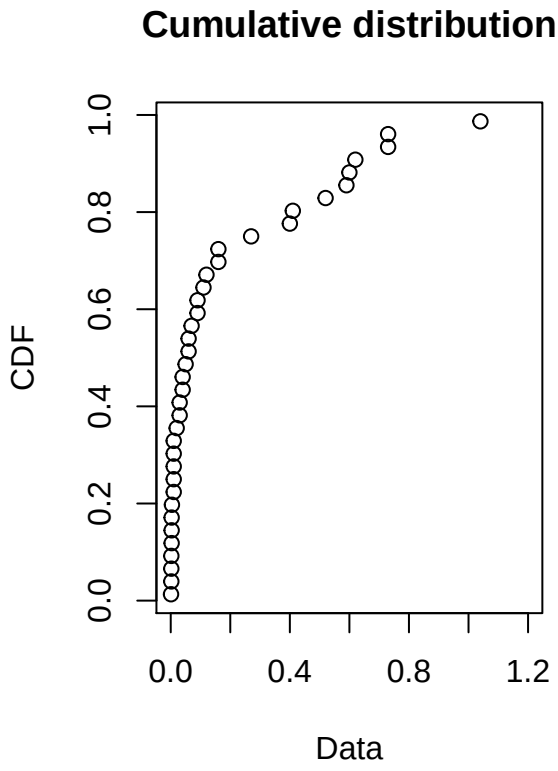

```
plotdist(D6_3$rain)
```





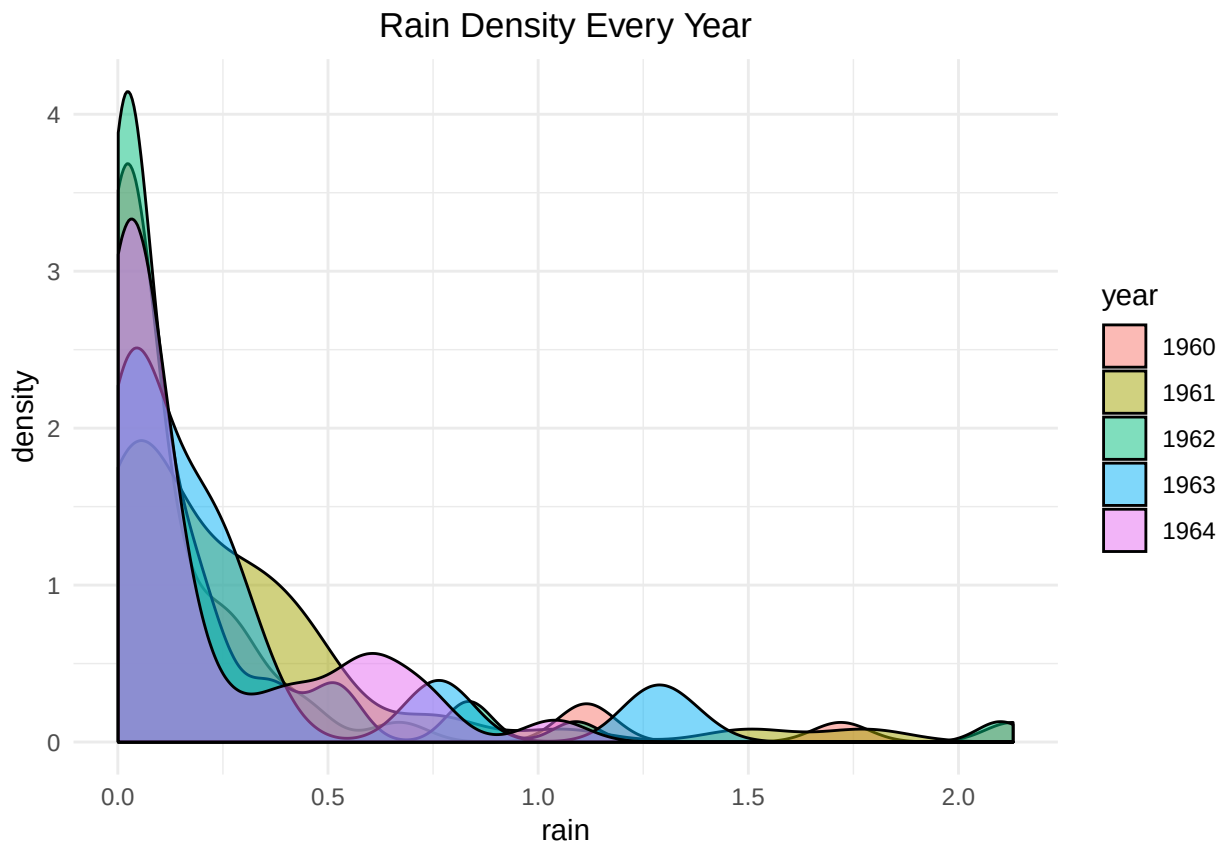

```
plotdist(D6_4$rain)
```



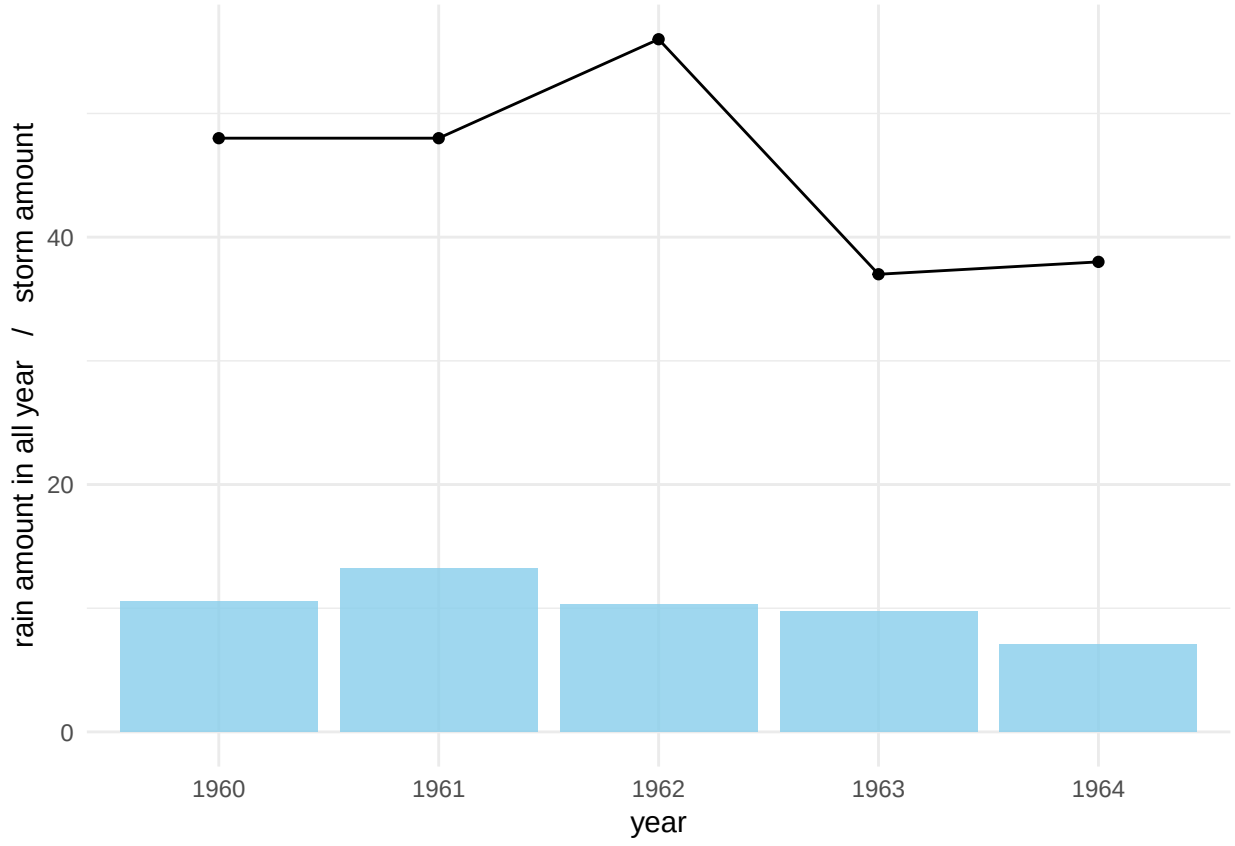



```
d1 <- D6 %>%
  group_by(year) %>%
  summarise(all_rain=sum(rain))
d1$storm_num <- c(48,48,56,37,38)
d1 <- d1 %>% mutate(mean_rain=round(all_rain/storm_num,3))

ggplot()+
  geom_density(data=D6,mapping=aes(rain,fill=year),alpha=0.5)+
  theme_minimal()+
  ggtitle("Rain Density Every Year")+
  theme(plot.title = element_text(hjust = 0.5))
```



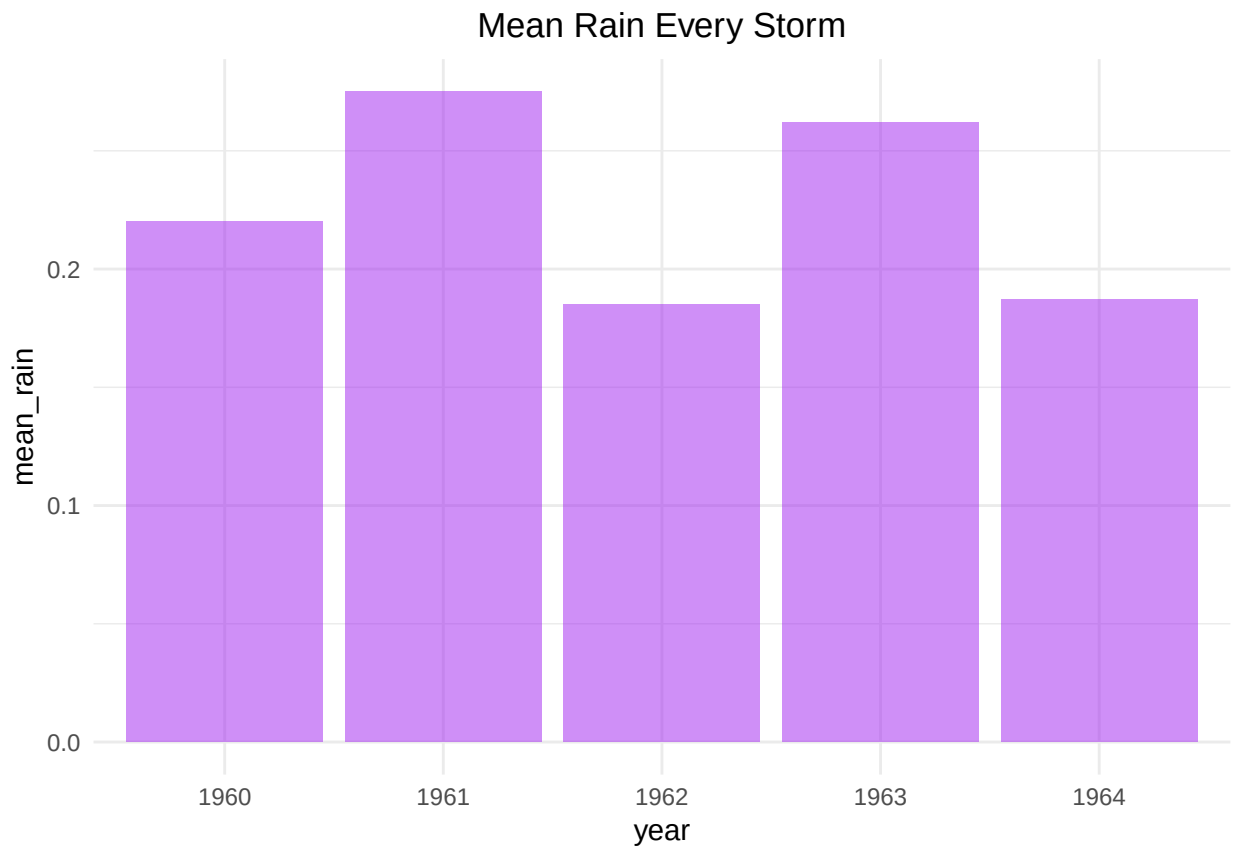

```
ggplot(d1)+
  geom_bar(mapping=aes(x=year,y=all_rain),fill="skyblue",alpha=.8,stat="identity")+
  geom_line(mapping=aes(x=year,y=storm_num,group=1))+
  geom_point(mapping=aes(x=year,y=storm_num))+
  theme_minimal()+
  ylab("rain amount in all year   /   storm amount")
```




```
  ggtitle("All Rain Every Year")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## NULL
```

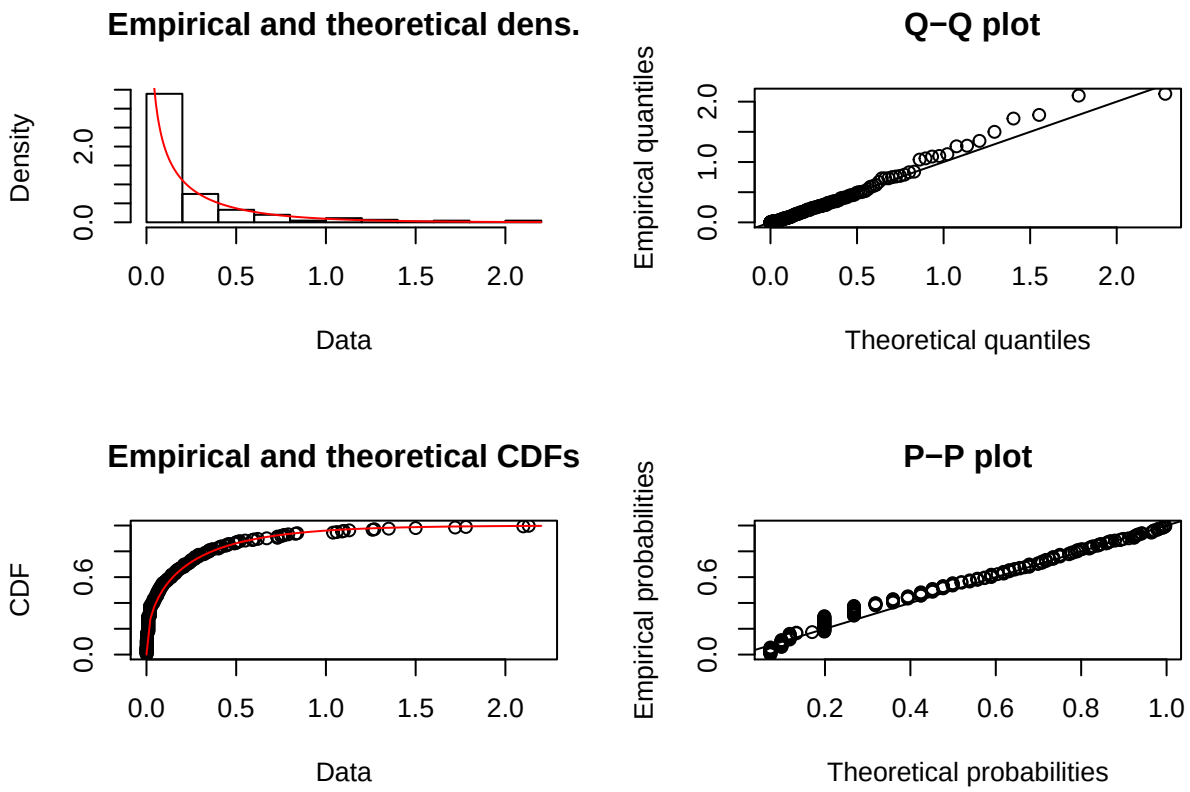
```
ggplot(d1)+
  geom_bar(mapping=aes(x=year,y=mean_rain),fill="purple",alpha=.5,stat="identity")+
  ggtitle("Mean Rain Every Storm")+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```



Distribution plots and the density plots indicate that the 5 years are similar. The first barplot shows that 1961 was wetter than the other 4 years, and it is because storms produced more rain since the storm amount in 1961 is not and highest and the mean rain every storm in 1961 is the highest according to the last barplot.

2. Is gamma distribution a goot fit for the data?

```
#Test whether it is gamma distribution
if_gamma <- fitdist(D6$rain, "gamma")
plot(if_gamma)
```



Both QQ-plot and P-P plot show that gamma distribution is a great fit of the rain data. Therefore, I totally agree with Changnon and Huff. ## what they might consider

3. estimate

```
set.seed(1234)

MoM <- fitdist(D6$rain, "gamma",method = "mme")
MoM_Boot <- bootdist(MoM)
summary(MoM_Boot)
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.3916308 0.2764478 0.5228296
## rate  1.7380433 1.1496196 2.4796917
```

```
MLE <- fitdist(D6$rain, "gamma",method = "mle")
MLE_Boot <- bootdist(MLE)
summary(MLE_Boot)
```

```
## Parametric bootstrap medians and 95% percentile CI
##          Median      2.5%     97.5%
## shape 0.4428035 0.3860241 0.5163245
## rate  1.9829288 1.5704146 2.5297209
```

The result tables above show that:

1. For Method of Moments: Estimate for shape parameter is 0.39, and the 95% confidence interval is (0.28, 0.52).

Estimate for rate parameter is 1.74, and the 95% confidence interval is (1.15, 2.48).

2. For MLE methods: Estimate for shape parameter is 0.44, and the 95% confidence interval is (0.39, 0.52).

Estimate for rate parameter is 1.98, and the 95% confidence interval is (1.57, 2.53).

Therefore, the variances of MLE method of two estimates are both narrower than vairance of MoM estimates. As a result, I would present MLE method to give the estimates.

Analysis of decision theory article

1. Derive (10a), (10b), (10c) (10a) (10b) (10c) are the posterior mean for β.

Prior distribution for β is Beta(c,d), where the density function is:

$$f(x) = \frac{x^{c-1}(1-x)^{d-1}}{B(c,d)}$$

The Binomial likelihood is:

$$p^n(1-p)^{N-n}$$

Therefore, the posterior density funtion is Beta(c+n, d+N-n):

$$p(x) = \frac{x^{c+n-1}(1-x)^{N-n+d-1}}{B(c+n, d+N-n)}$$

Thus, the mean is:

$$\begin{aligned}
\delta(n) = 0 \quad & for \frac{c+n}{c+d+N} < \alpha \\
\delta(n) = \lambda \quad & for \frac{c+n}{c+d+N} = \alpha, where\, 0 \leq \lambda \leq 1 \\
\delta(n) = 1 \quad & for \frac{c+n}{c+d+N} > \alpha
\end{aligned}$$

2. Reproduce 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)

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"))[1:55,]
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)
    delt = delta(n0,lambda,i)
    sum = sum+ f*delt
  }
  return(sum)
}
```

Becky helped me with this problem.