

677 Final Project

Yifeng Luo

1 Statistics and the Law Arguement that the difference between the rates of mortgage application refusals of white applicants and minority applicatns constituted evidence of discrimination. Test (1) the data are sufficient evidence of discrimination to warrant corrective action and (2) the data are not sufficient.

H0: mortgage refusal rate of white applicants are the same as that of minority applicant H1: mortgage refusal rate of white applicants are lower than that of minority applicant

```
#Store the data
Min<-c(20.90,23.23,23.10,30.40,42.70,62.20,39.5,38.40,26.20,55.90,49.70,44.60,36.40,32.00,10.60,34.30,42.70)
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)
data1 <- data.frame(
  group = rep(c("Min", "White"), each = 20),
  percent = c(Min, White)
)

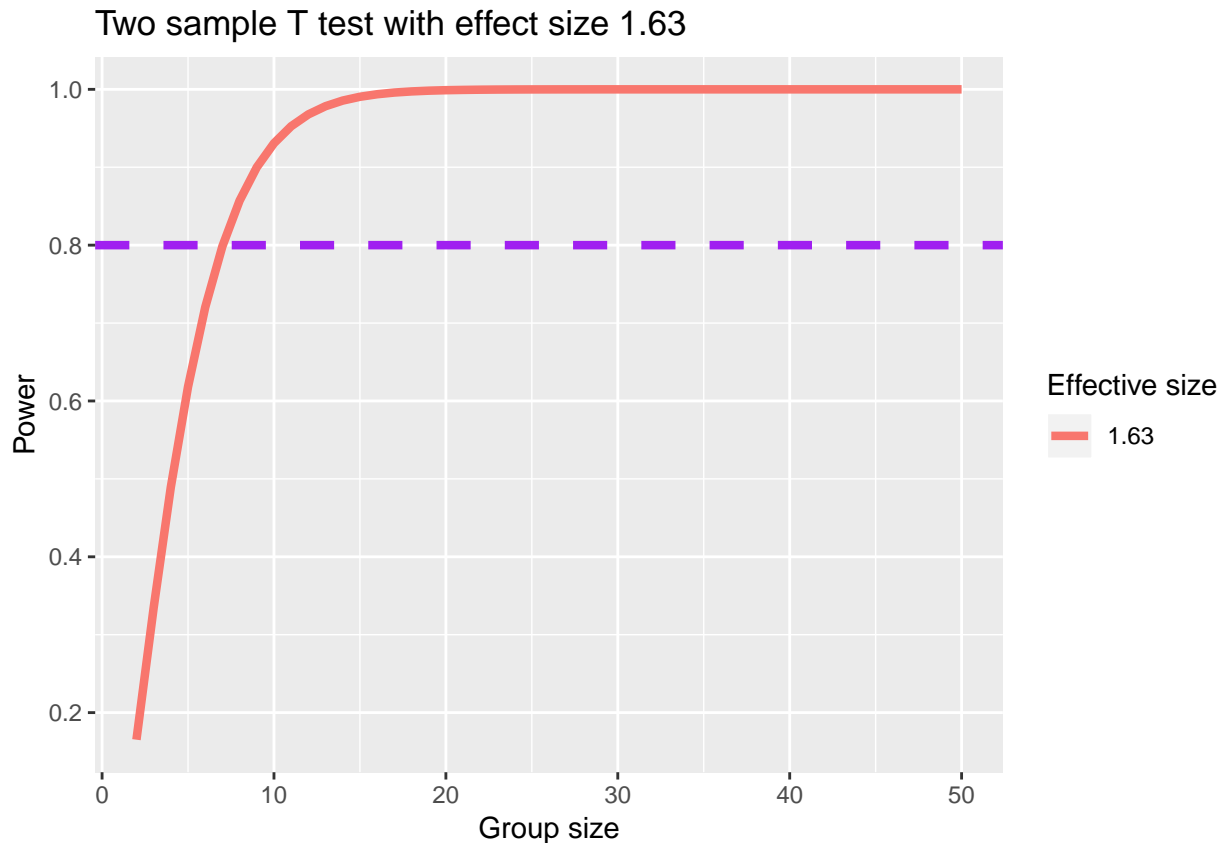
# Two sample T-test to test whether the refusal rate for minority applicants is greater than the refusal rate for white applicants
res<-t.test(percent ~ group, data = data1,
  var.equal = TRUE, alternative = "greater")
res$p.value
```

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

The p-value of this t test is 1.279668e-07, which is less than the significant level $\alpha = 0.05$. We

```
# 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) +
  ylab("Power") + scale_y_continuous(breaks = seq(0, 1, by = 0.2)) + ggtitle("Two sample T test with ef:
```



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?

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
```

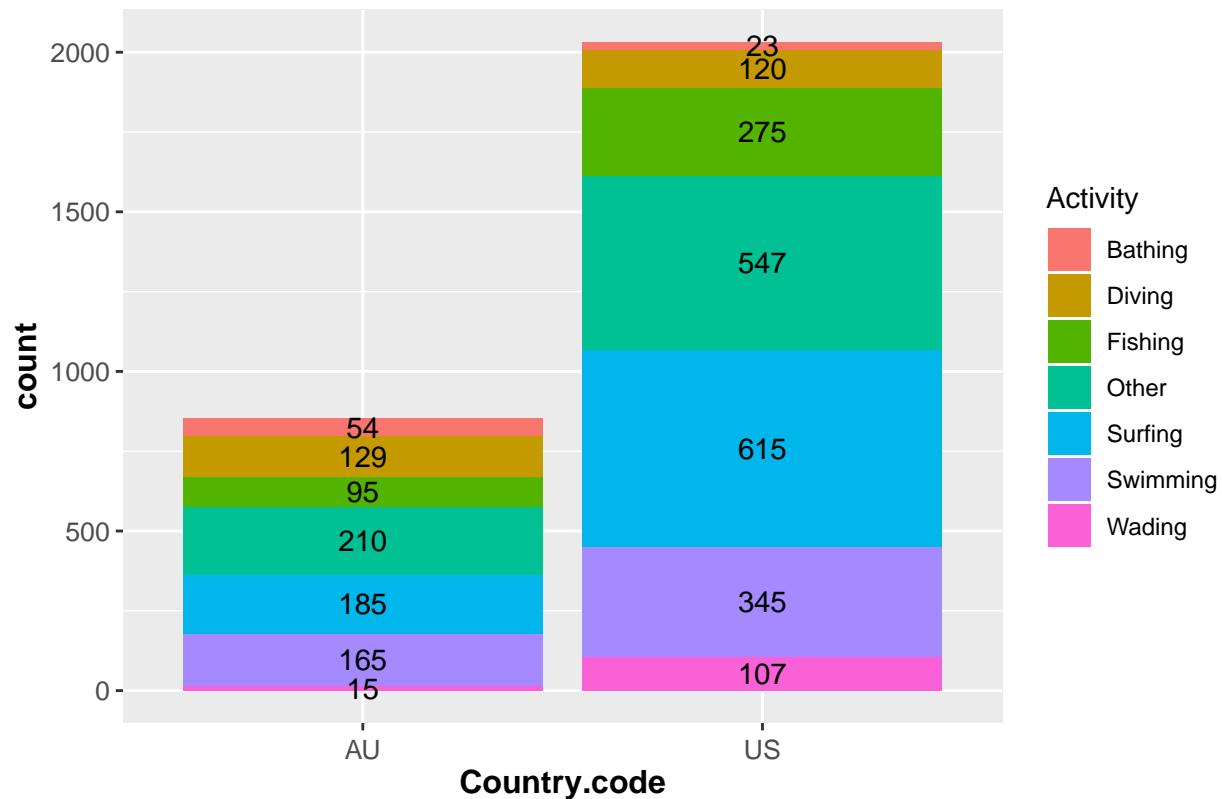
3 How deadly are sharks?

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 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 about
data3<-data3%>%filter(Country.code=="US"|Country.code=="AU"&Type=="Unprovoked")
temp<-data3%>%group_by(Country.code,Activity)%>%summarise(count=n())%>% ungroup()%>%group_by(Country.code)
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(
    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)
```



```
# According to the bar plot and data frame, US sharks made more attacks in total and attacks in surfing

# Transfer dataframe into matrix 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","Wading")),
                byrow=T)
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 alpha=0.05. Therefore,
```

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't 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 $t_1 - t_2 = h$, which has equal detectability. 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

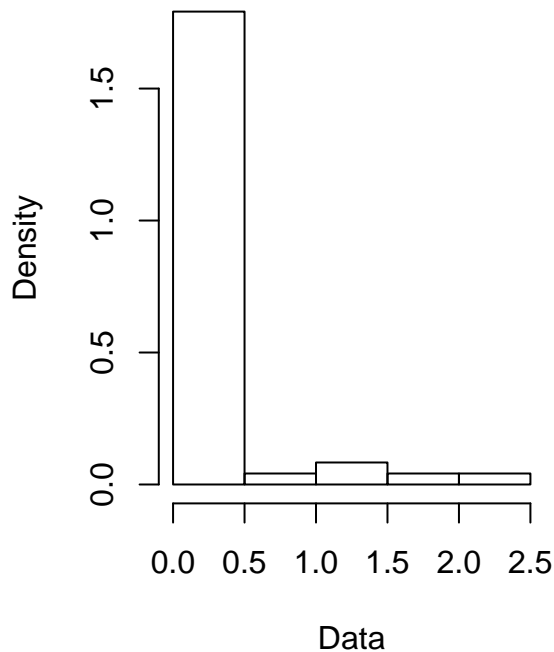
Case2

Case3

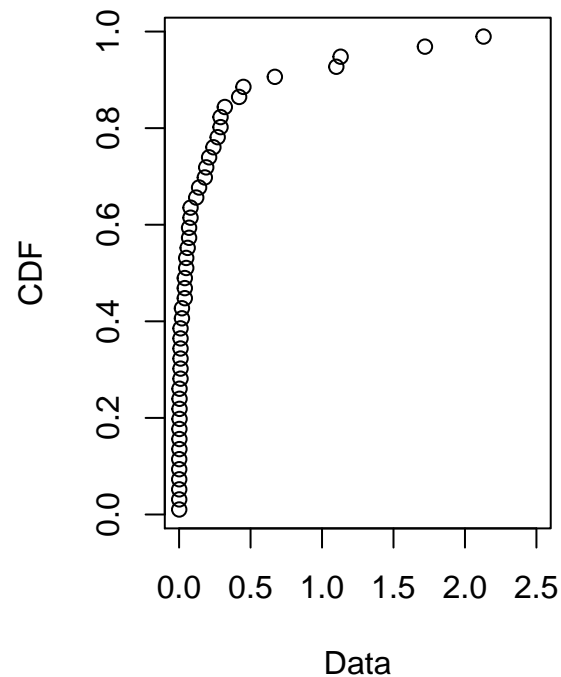
```
# read the data
data60 <- read.table("ill-60.txt", quote="\"", comment.char="")
data60 <- as.numeric(as.array(data60 [,1]))
data61 <- read.table("ill-61.txt", quote="\"", comment.char="")
data61 <- as.numeric(as.array(data61 [,1]))
data62 <- read.table("ill-62.txt", quote="\"", comment.char="")
data62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

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

Histogram

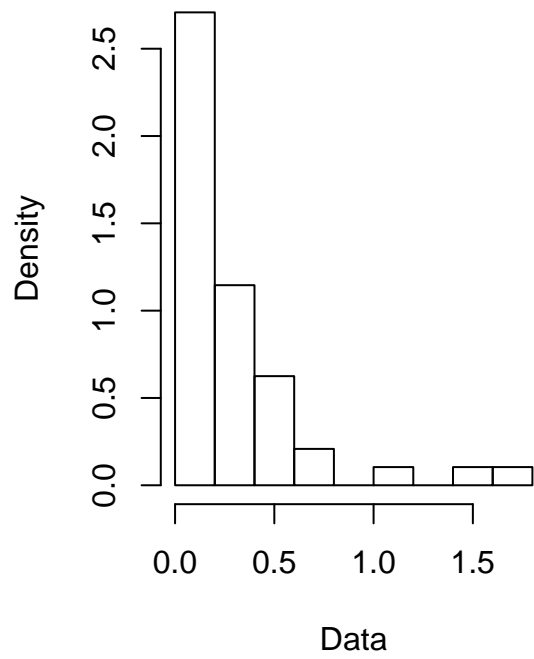


Cumulative distribution

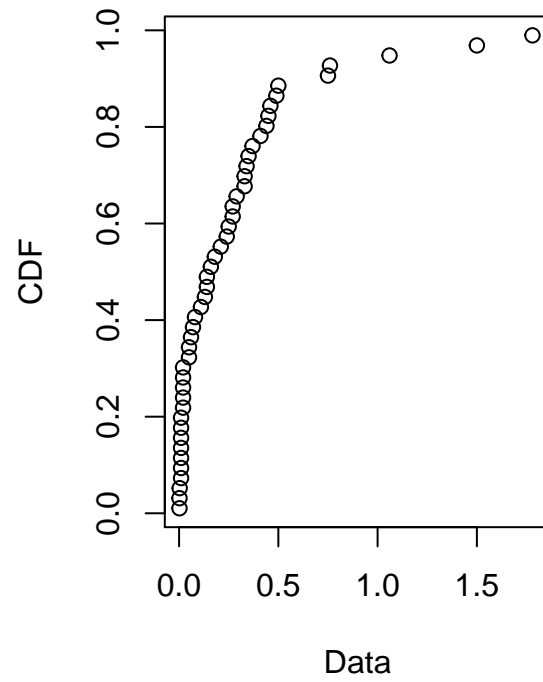


```
plotdist(data61)
```

Histogram

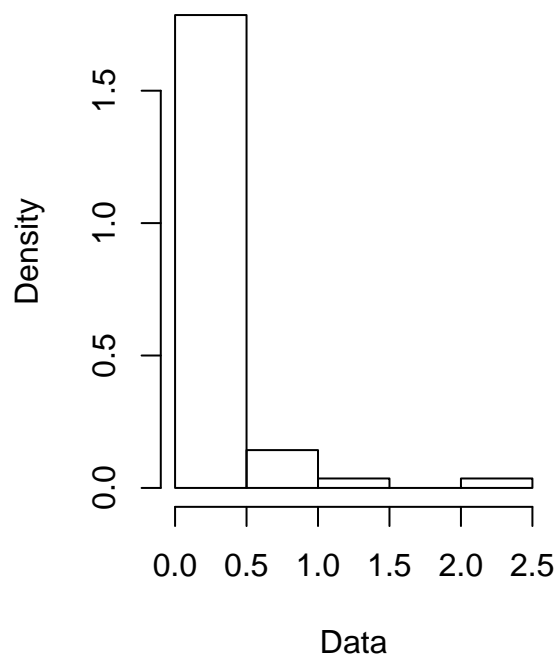


Cumulative distribution

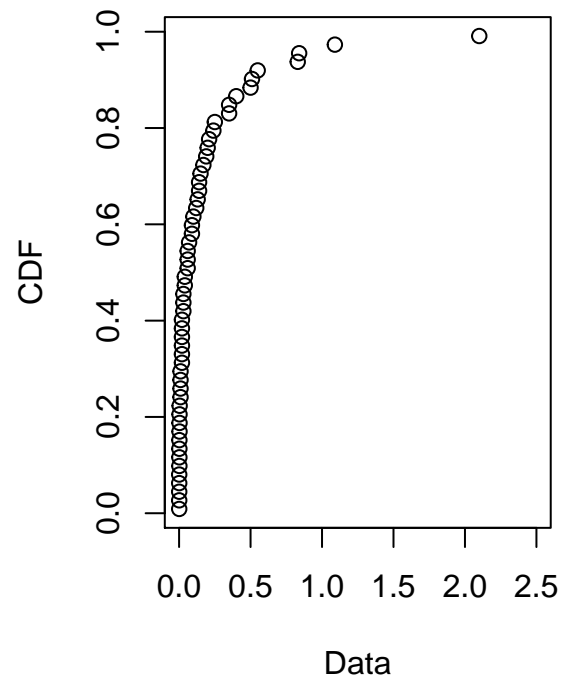


```
plotdist(data62)
```

Histogram

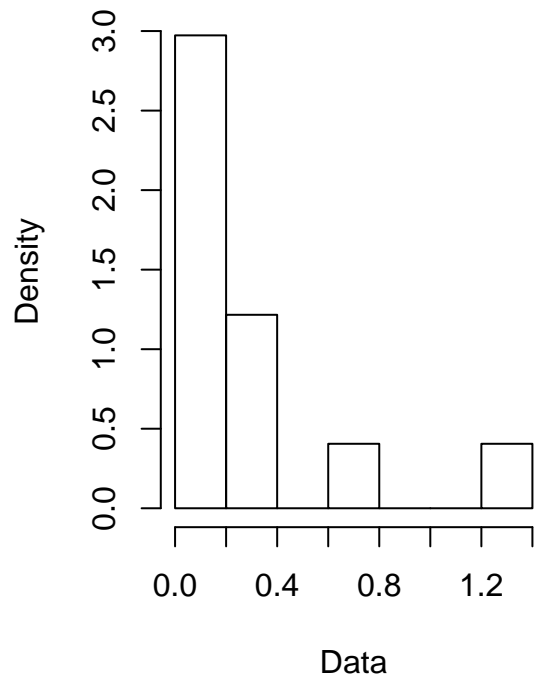


Cumulative distribution

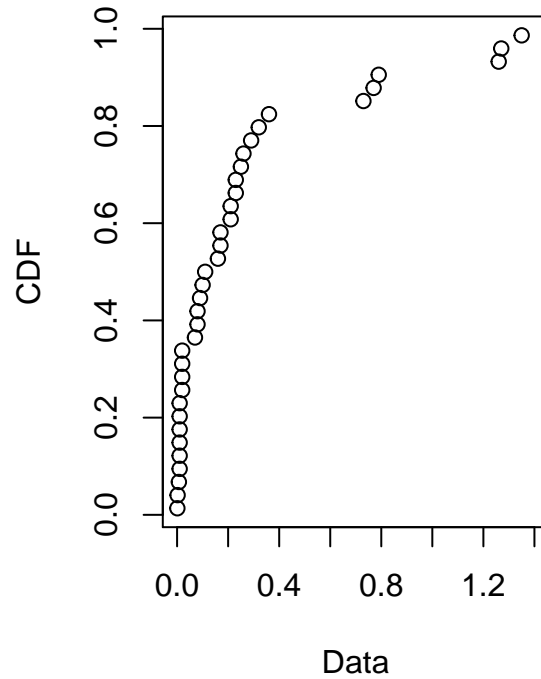


```
plotdist(data63)
```

Histogram

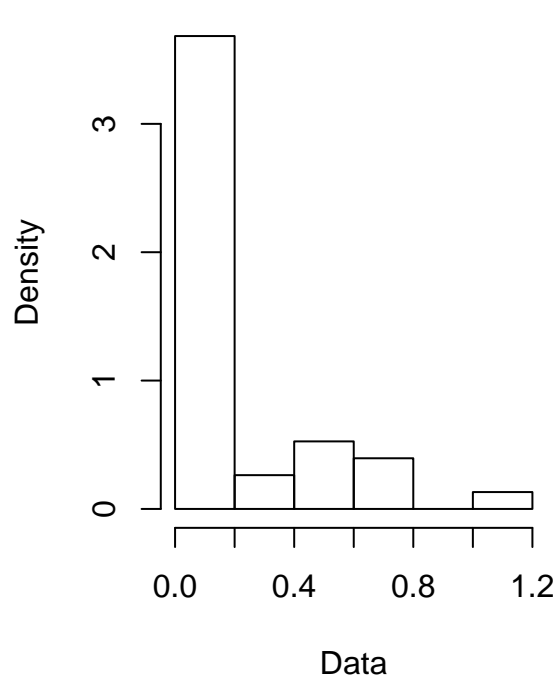


Cumulative distribution

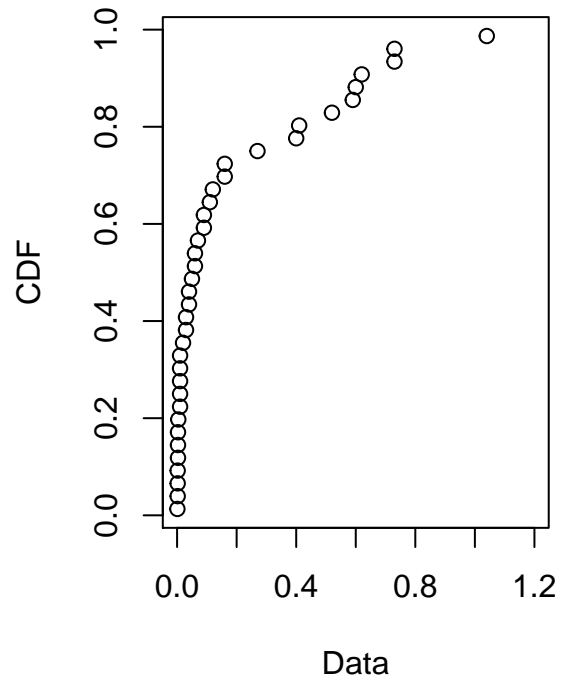


```
plotdist(data64)
```

Histogram



Cumulative distribution



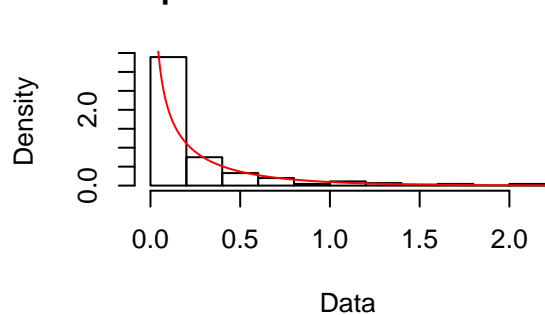
```
SumOfRain<-as.data.frame(t(c(sum(data60),sum(data61),sum(data62),sum(data63),sum(data64))))  
colnames(SumOfRain)[1:5]<-c("Total Rainfall in 1960","Total Rainfall in 1961","Total Rainfall in 1962",  
kable(SumOfRain)
```

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

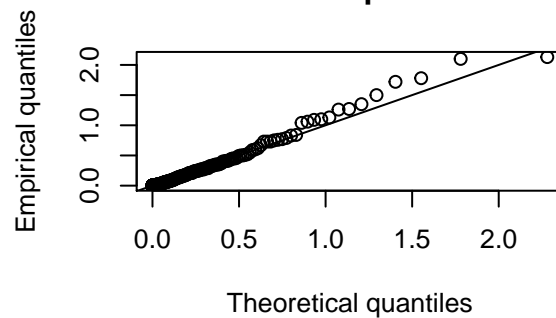
According to the distribution plot, five years are similar. 1961 is more wetter than others since it

```
#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)
```

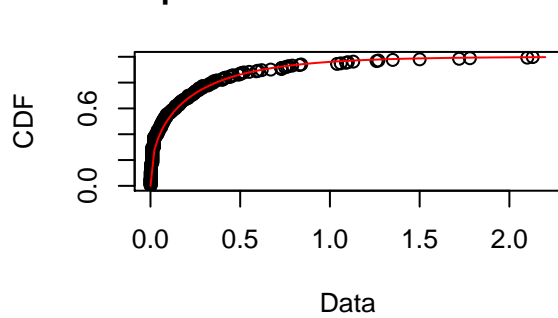
Empirical and theoretical dens.



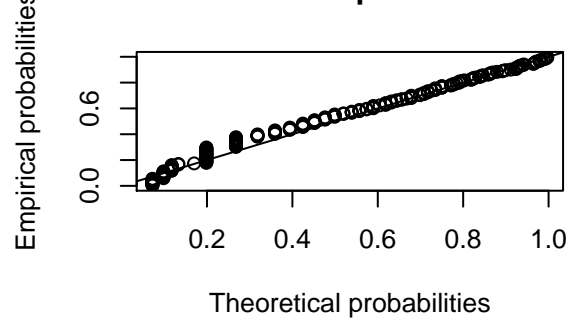
Q-Q plot



Empirical and theoretical CDFs



P-P plot



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

```
# 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 0.3943445 0.2795922 0.5303857
## rate  1.7640324 1.1487165 2.5032988
```

```
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 0.4433798 0.3836208 0.5088105
## rate 1.9785916 1.5448273 2.5743222
```

#For method of moment the 95% confidence interval of shape from bootstrap sample is (0.27,0.53), the ra

Analysis of decision theory article