

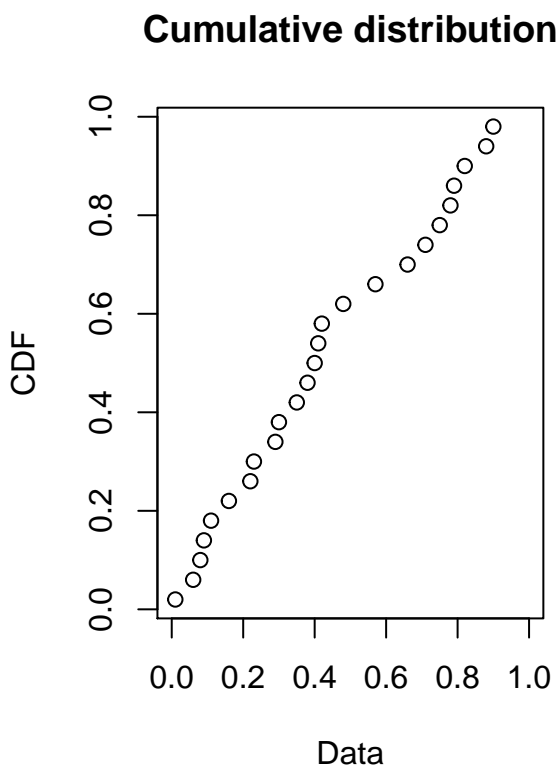
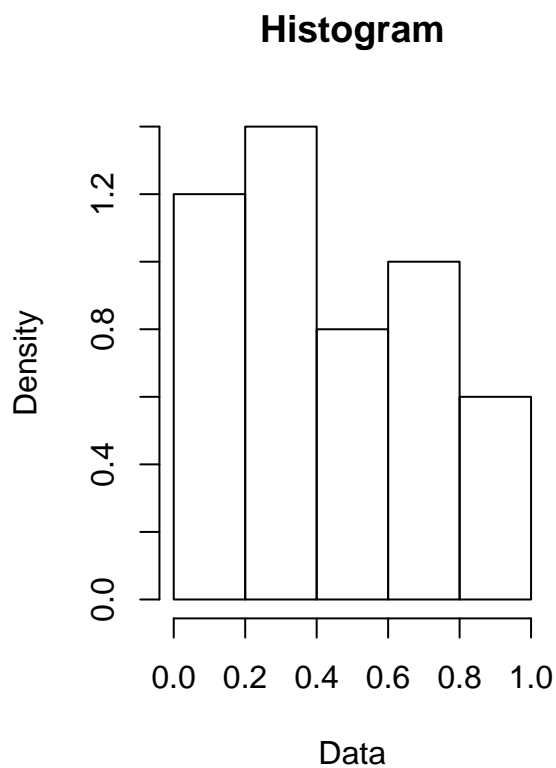
HW_EmpiricalDistribution

Tingrui Huang

March 4, 2019

Problem 1

```
library(readr)
library(tidyverse)
library(fitdistrplus)
# Question a
maybe_uniform <- read_table2("../ma677/data/maybe_uniform.txt", col_names = FALSE)
maybe_uniform <- unlist(maybe_uniform)
plotdist(maybe_uniform)
```



```
ks.test(maybe_uniform, punif(25, 0, 1))

##
## Two-sample Kolmogorov-Smirnov test
##
## data: maybe_uniform and punif(25, 0, 1)
## D = 1, p-value = 0.07692
## alternative hypothesis: two-sided
```

```

# P-value is 0.076 indicates we can not reject the H0, therefore, we might say the empirical distribution is normal

# Question b
testmodel <- function(x){
  return(2/3*I(0<x & x<=1/2)+1/2*I(1/2<x & x<1))
}
ks.test(testmodel(seq(-2,2,by=0.01)),"punif")

## Warning in ks.test(testmodel(seq(-2, 2, by = 0.01)), "punif"): ties should
## not be present for the Kolmogorov-Smirnov test

##
## One-sample Kolmogorov-Smirnov test
##
## data: testmodel(seq(-2, 2, by = 0.01))
## D = 0.75312, p-value < 2.2e-16
## alternative hypothesis: two-sided

# The p-value is smaller than 0.05 indicates we can reject the null hypothesis, hence we can say the tw

```

Problem 2

```

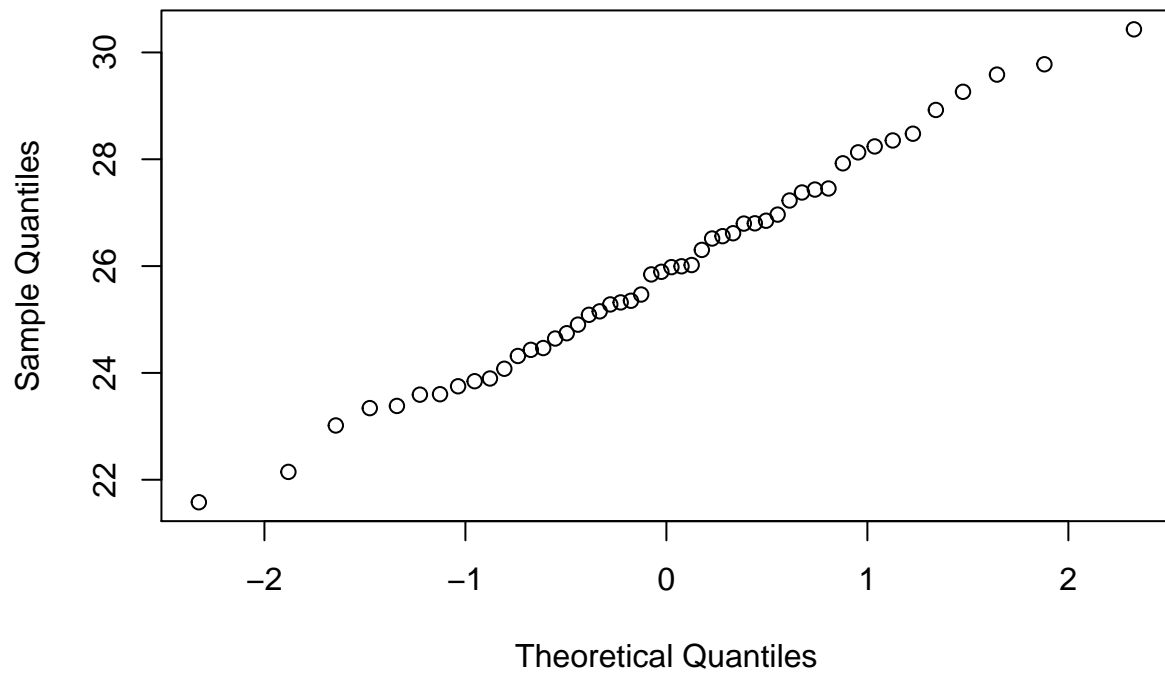
maybe_normal <- read_table2("../ma677/data/maybe_normal.txt", col_names = FALSE)
maybe_normal <- unlist(maybe_normal)
ks.test(maybe_normal,pnorm(50,26,4))

##
## Two-sample Kolmogorov-Smirnov test
##
## data: maybe_normal and pnorm(50, 26, 4)
## D = 1, p-value = 0.03922
## alternative hypothesis: two-sided

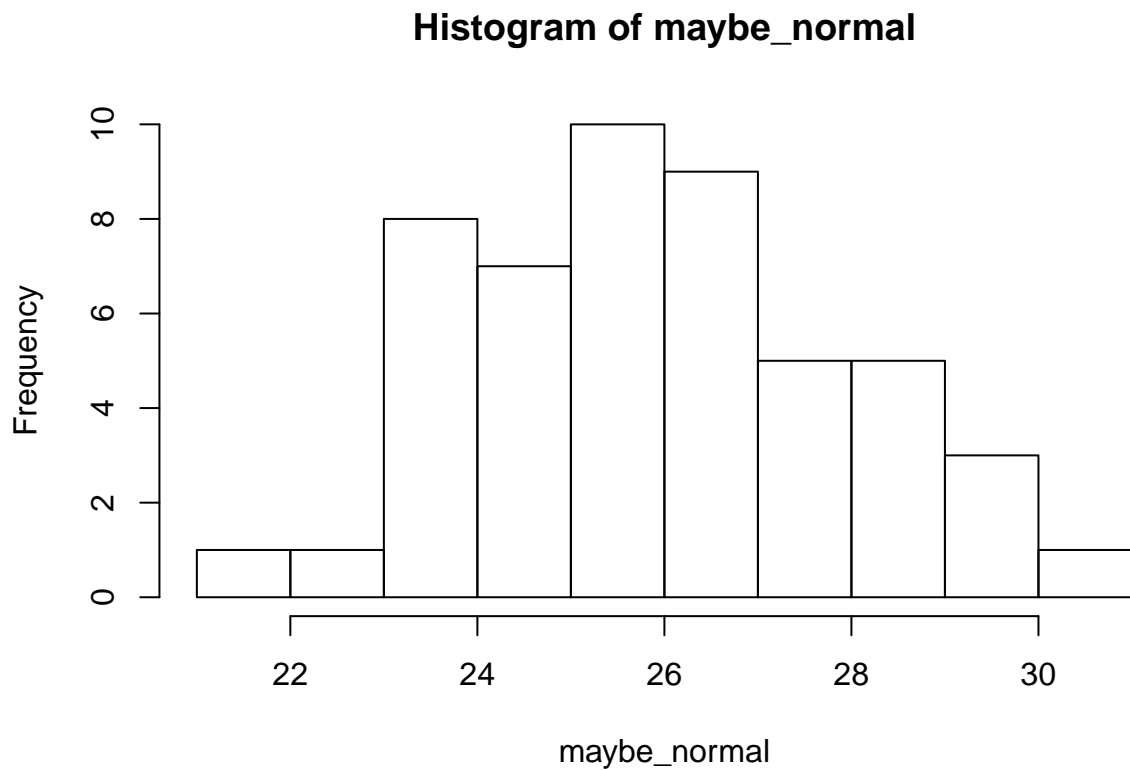
# P-value is 0.039 indicates we reject the H0, therefore, we might say the empirical distribution is di.
qqnorm(maybe_normal)

```

Normal Q-Q Plot



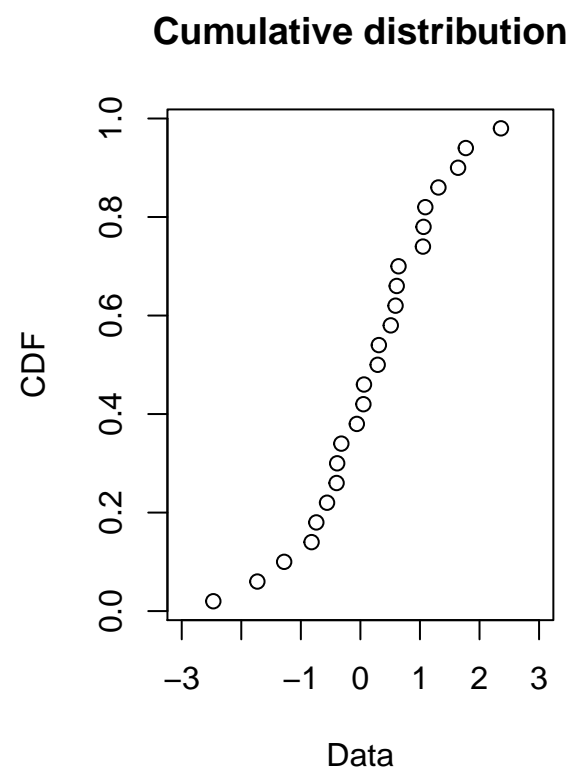
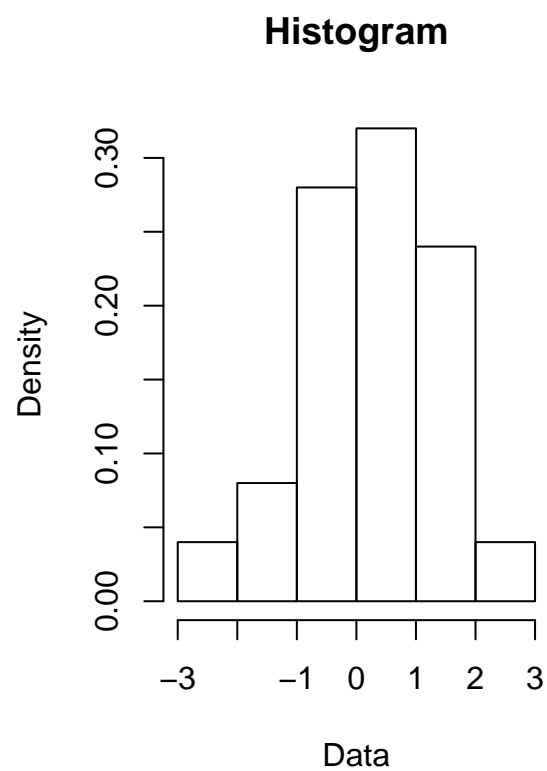
```
hist(maybe_normal)
```



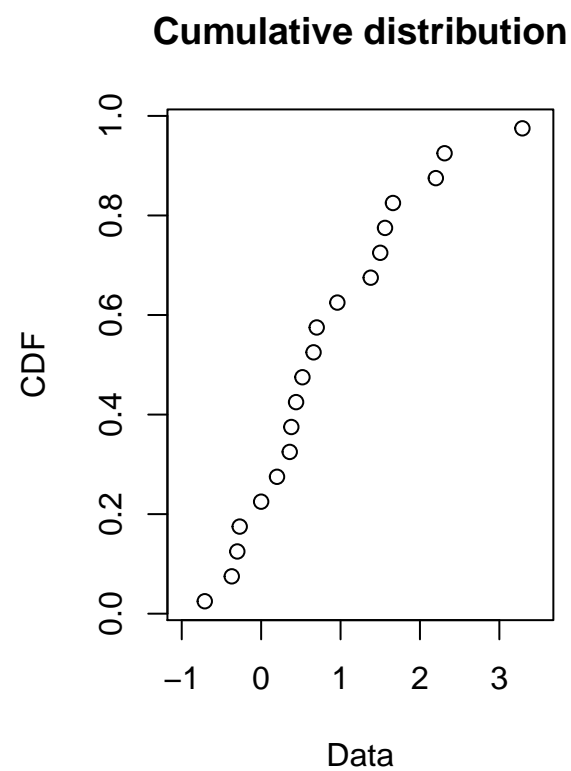
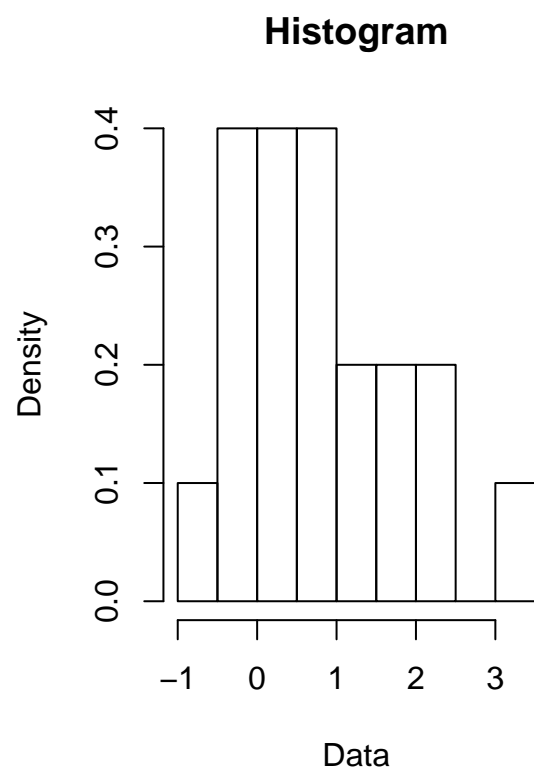
From the qnorm plot and histogram we can tell the empirical distribution is close to normal distribut

Problem 3

```
maybe_same_1 <- read_table2("../ma677/data/maybe_same_1.txt", col_names = FALSE)
maybe_same_2 <- read_table2("../ma677/data/maybe_same_2.txt", col_names = FALSE)
maybe_same_1 <- unlist(maybe_same_1)
maybe_same_2 <- unlist(maybe_same_2)
plotdist(maybe_same_1)
```

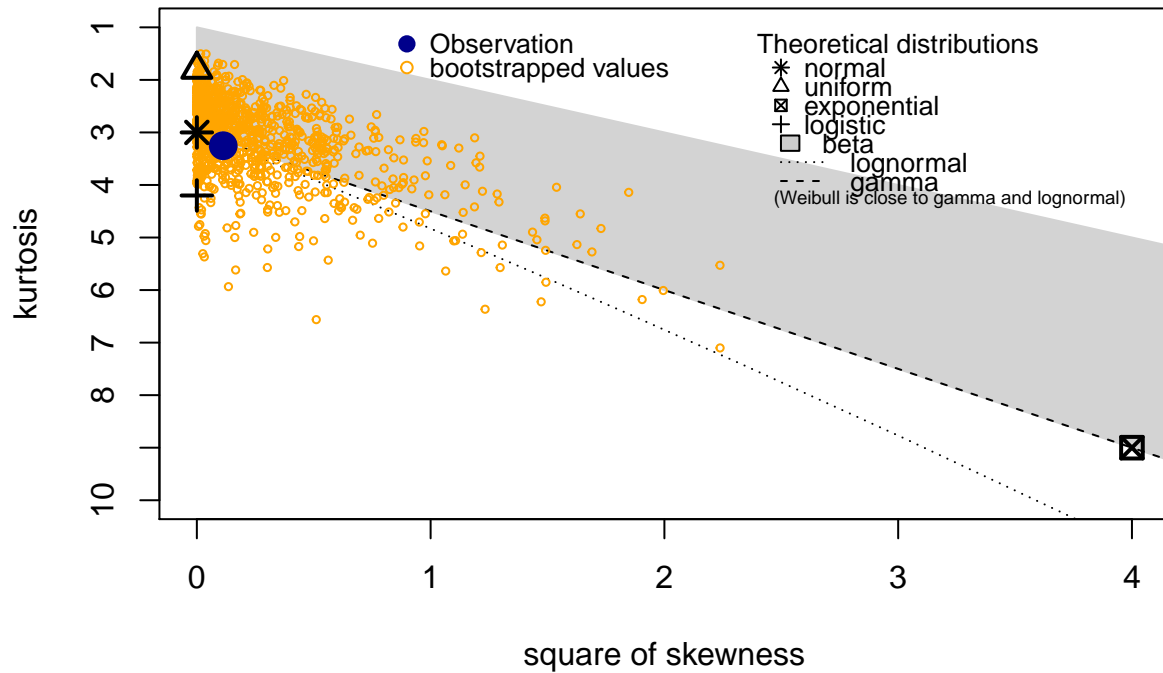


```
plotdist(maybe_same_2)
```



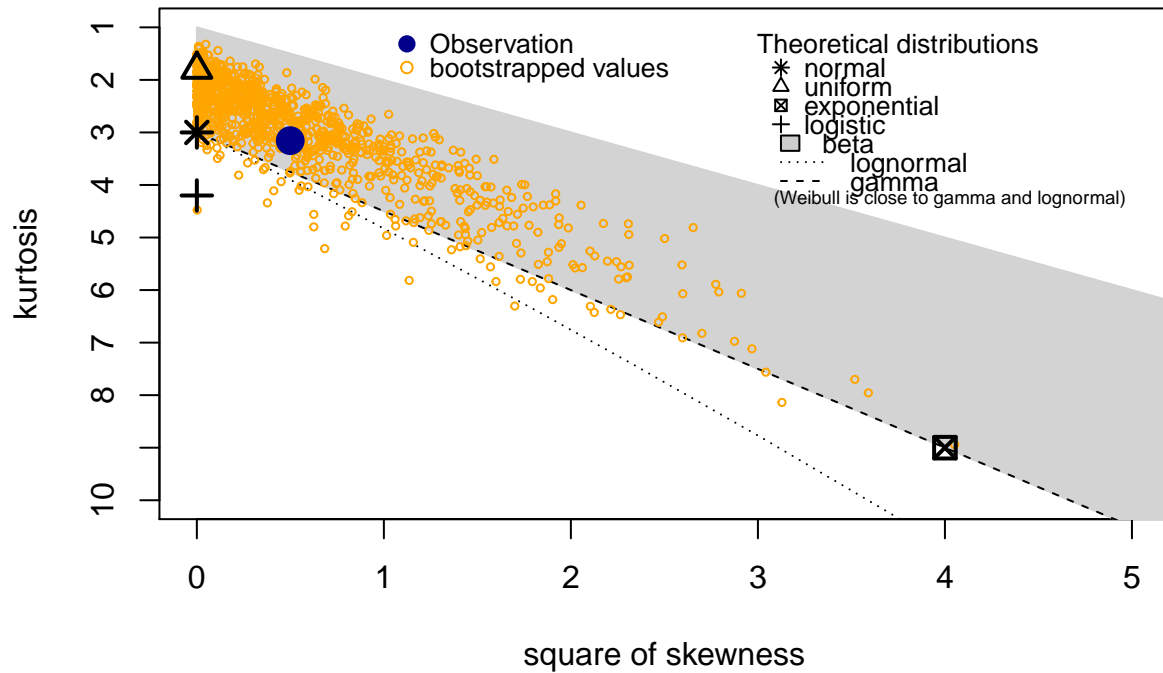
```
descdist(maybe_same_1, boot = 1000)
```

Cullen and Frey graph



```
## summary statistics
## -----
## min: -2.47   max:  2.36
## median:  0.29
## mean:  0.1828
## estimated sd:  1.117048
## estimated skewness: -0.3357935
## estimated kurtosis:  3.253393
descdist(maybe_same_2, boot = 1000)
```

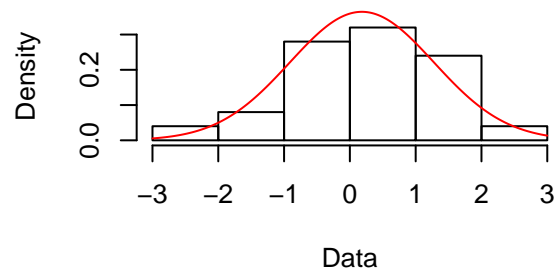
Cullen and Frey graph



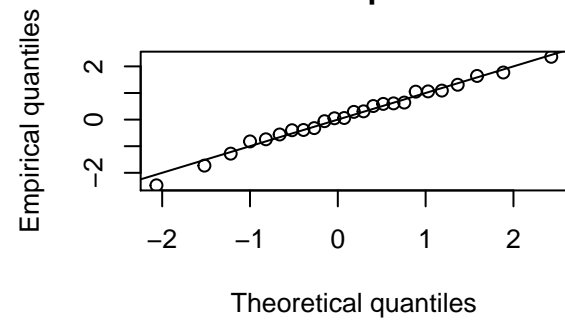
```
## summary statistics
## -----
## min: -0.71   max:  3.29
## median:  0.59
## mean:  0.8235
## estimated sd:  1.031347
## estimated skewness:  0.7067077
## estimated kurtosis:  3.156498
```

```
same1 <- fitdistr(maybe_same_1,distr = "norm")
same2 <- fitdistr(maybe_same_2,distr = "norm")
par(mfrow=c(1,2))
plot(same1)
```

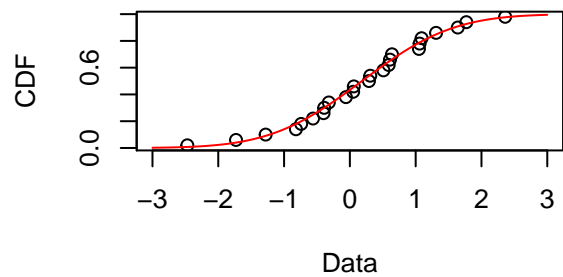

Empirical and theoretical dens.



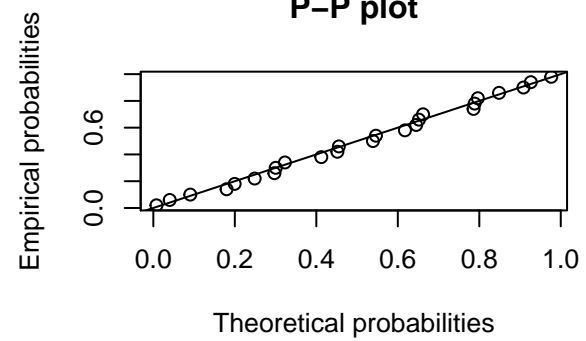
Q-Q plot



Empirical and theoretical CDFs

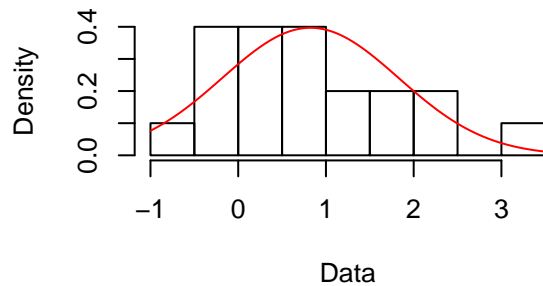


P-P plot

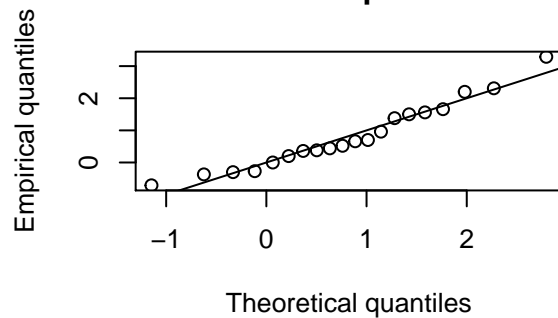


```
plot(same2)
```

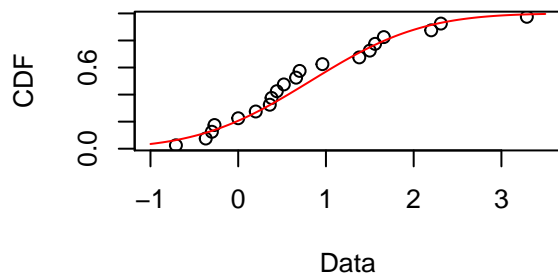
Empirical and theoretical dens.



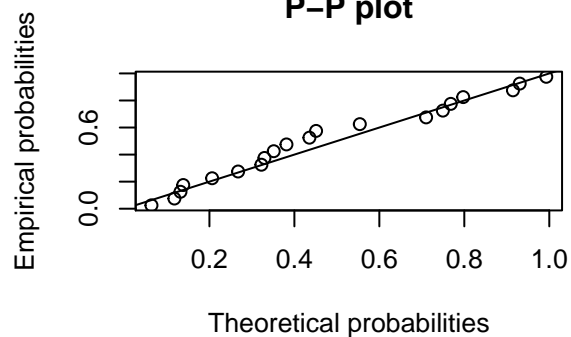
Q-Q plot



Empirical and theoretical CDFs



P-P plot



```
ks.test(maybe_same_1,maybe_same_2)
```

```
##  
## Two-sample Kolmogorov-Smirnov test  
##  
## data: maybe_same_1 and maybe_same_2  
## D = 0.27, p-value = 0.3357  
## alternative hypothesis: two-sided
```

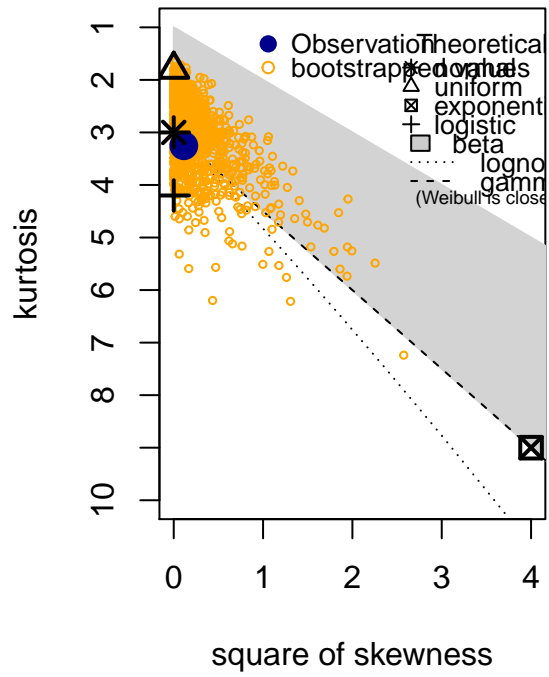
It seems both datasets are not from different distribution. The p-value 0.33 indicates we can not reject the null hypothesis.

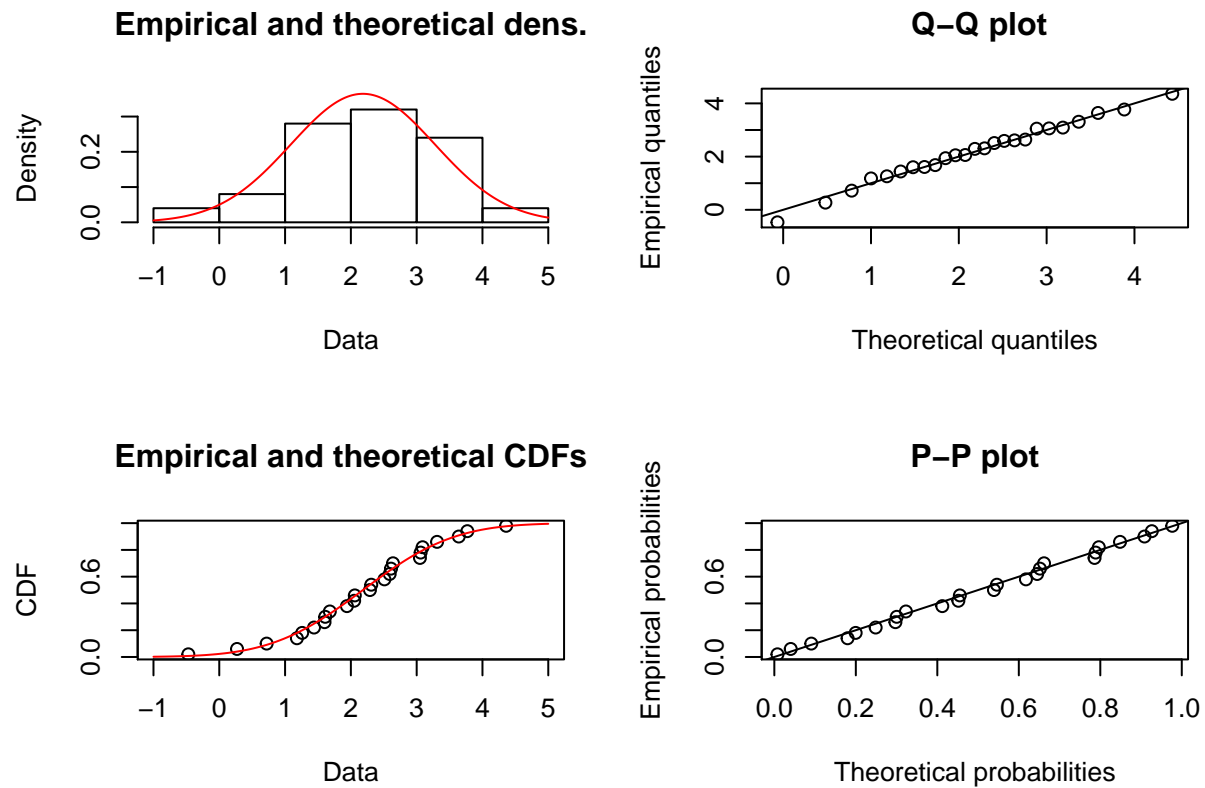
```
maybesame1.2 <- maybe_same_1+2  
descdist(maybesame1.2, boot = 1000)
```

```
## summary statistics  
## -----  
## min: -0.47 max: 4.36  
## median: 2.29  
## mean: 2.1828  
## estimated sd: 1.117048  
## estimated skewness: -0.3357935  
## estimated kurtosis: 3.253393
```

```
same1.2 <- fitdist(maybesame1.2,distr = "norm")  
plot(same1.2)
```

Cullen and Frey graph





```
ks.test(maybesame1.2,maybe_same_2)
```

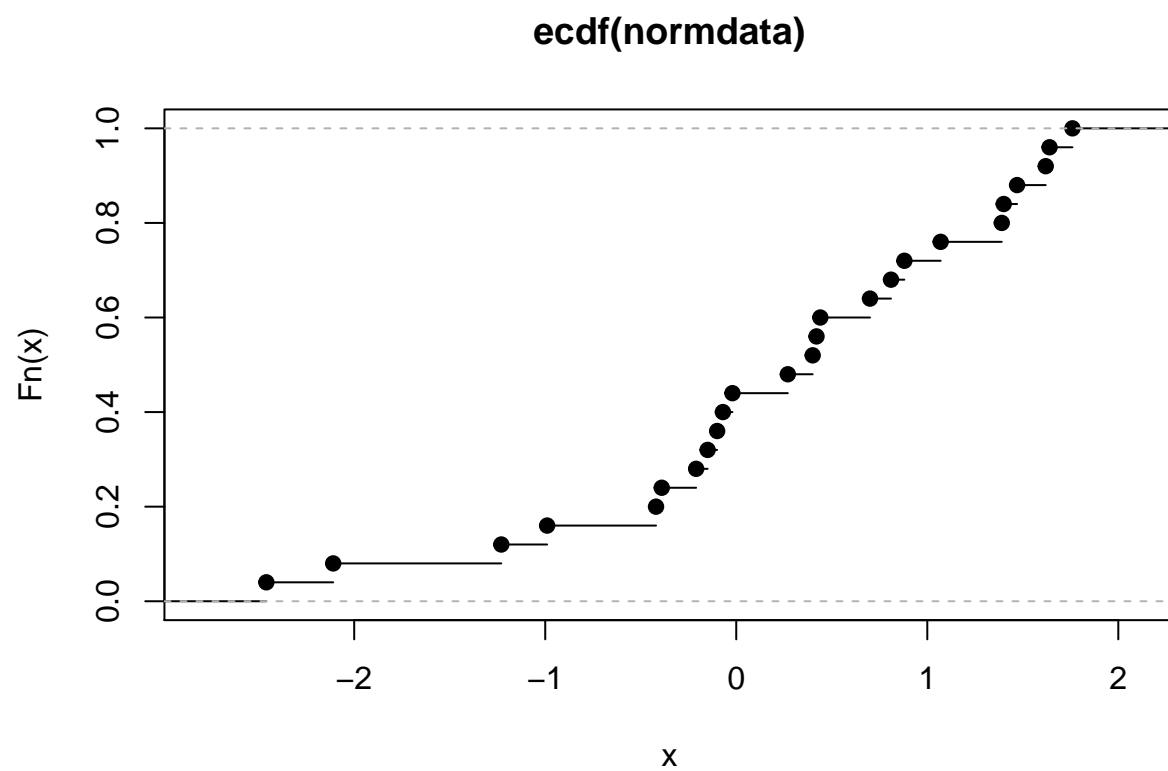
```
## Warning in ks.test(maybesame1.2, maybe_same_2): cannot compute exact p-  
## value with ties
```

```
##  
## Two-sample Kolmogorov-Smirnov test  
##  
## data: maybesame1.2 and maybe_same_2  
## D = 0.56, p-value = 0.001881  
## alternative hypothesis: two-sided
```

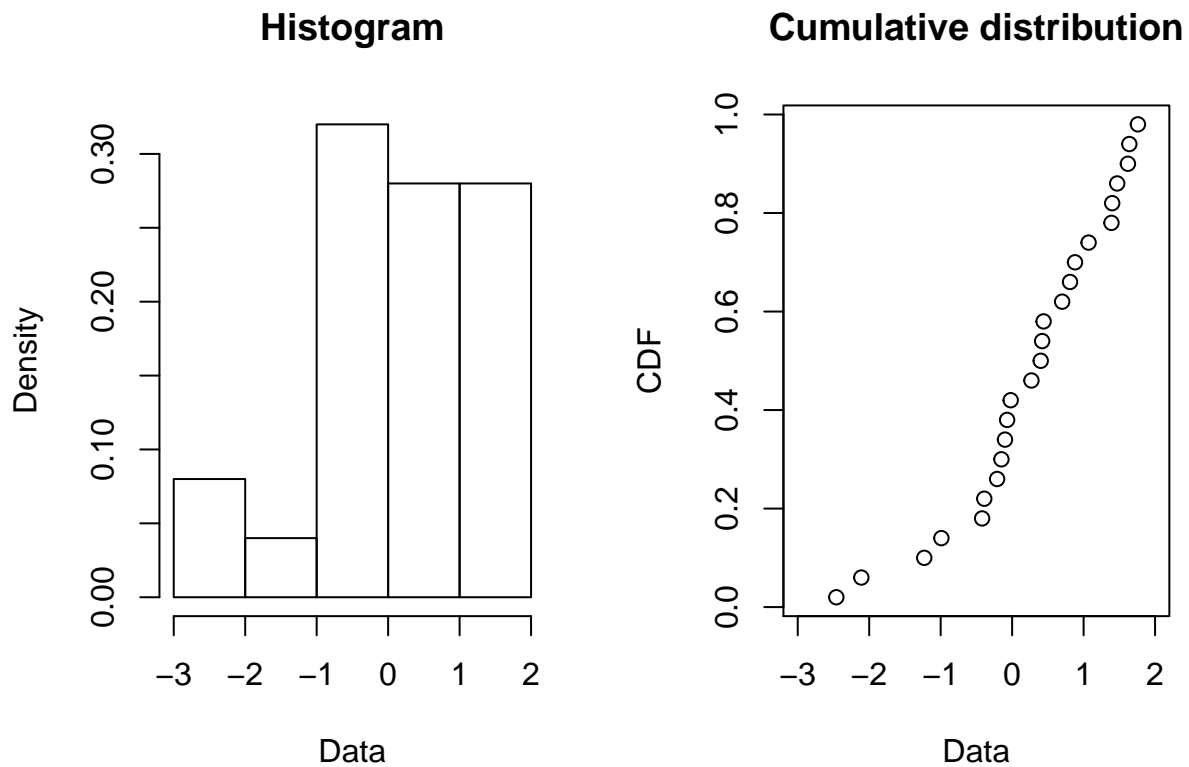
The p-value is very small and it indicates we can reject our null hypothesis, such that we can say th

Problem 4

```
normdata <- readRDS("../ma677/data/norm_sample.Rdata")  
plot(ecdf(normdata))
```



```
plotdist(normdata)
```



```
mean(normdata);var(normdata)
```

```
## [1] 0.2448
```

```
## [1] 1.251743
```

```
ks.test(normdata,"pnorm")
```

```
##
```

```
## One-sample Kolmogorov-Smirnov test
```

```
##
```

```
## data: normdata
```

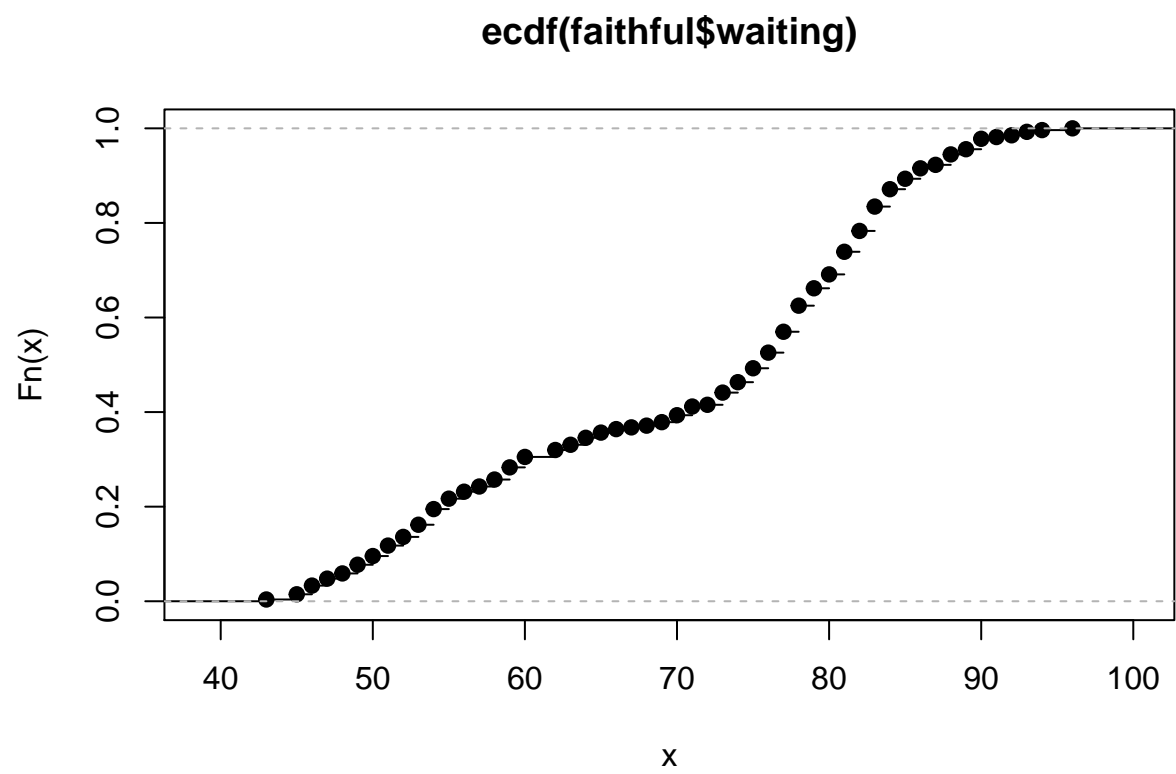
```
## D = 0.17724, p-value = 0.3683
```

```
## alternative hypothesis: two-sided
```

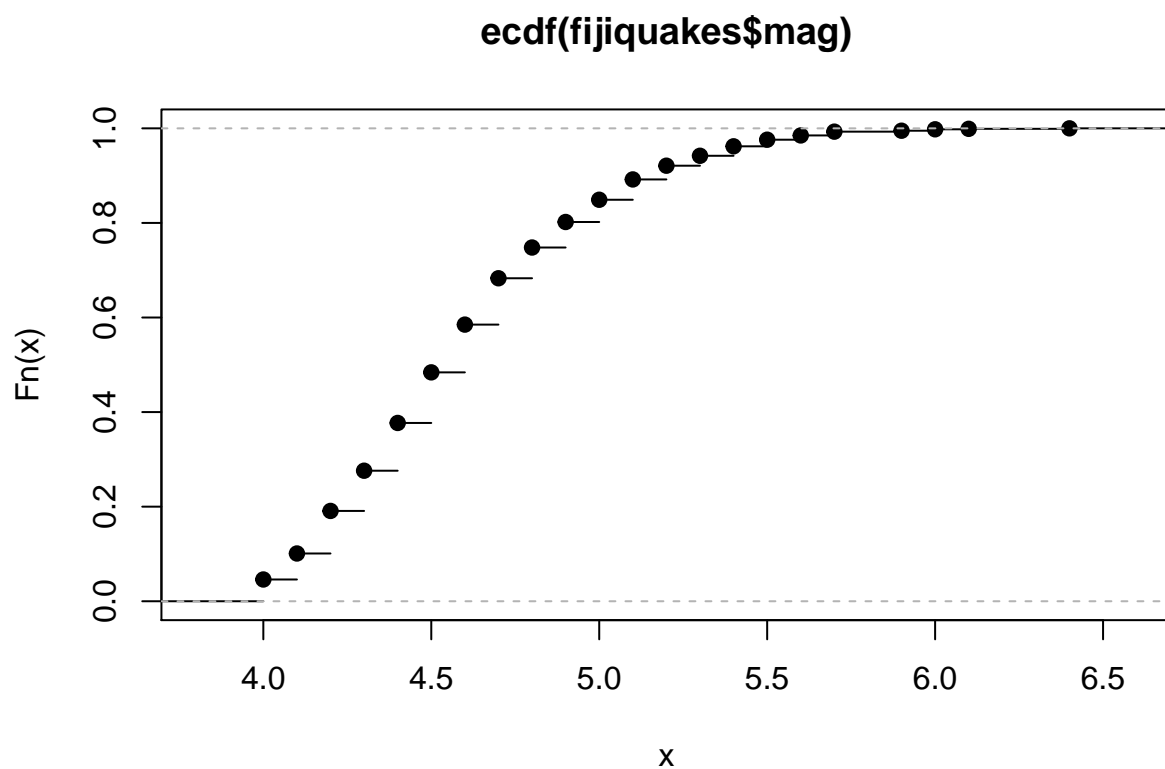
```
# P-value is 0.368 indicates the empirical distribution is the same normal distribution
```

Problem 5

```
faithful <- read.delim("../ma677/data/faithful.dat",header = TRUE,skip = 20,sep = "")
fijiquakes <- read.delim("../ma677/data/fijiquakes.dat",header = TRUE, sep = "")
# Plot ECDF
plot(ecdf(faithful$waiting))
```

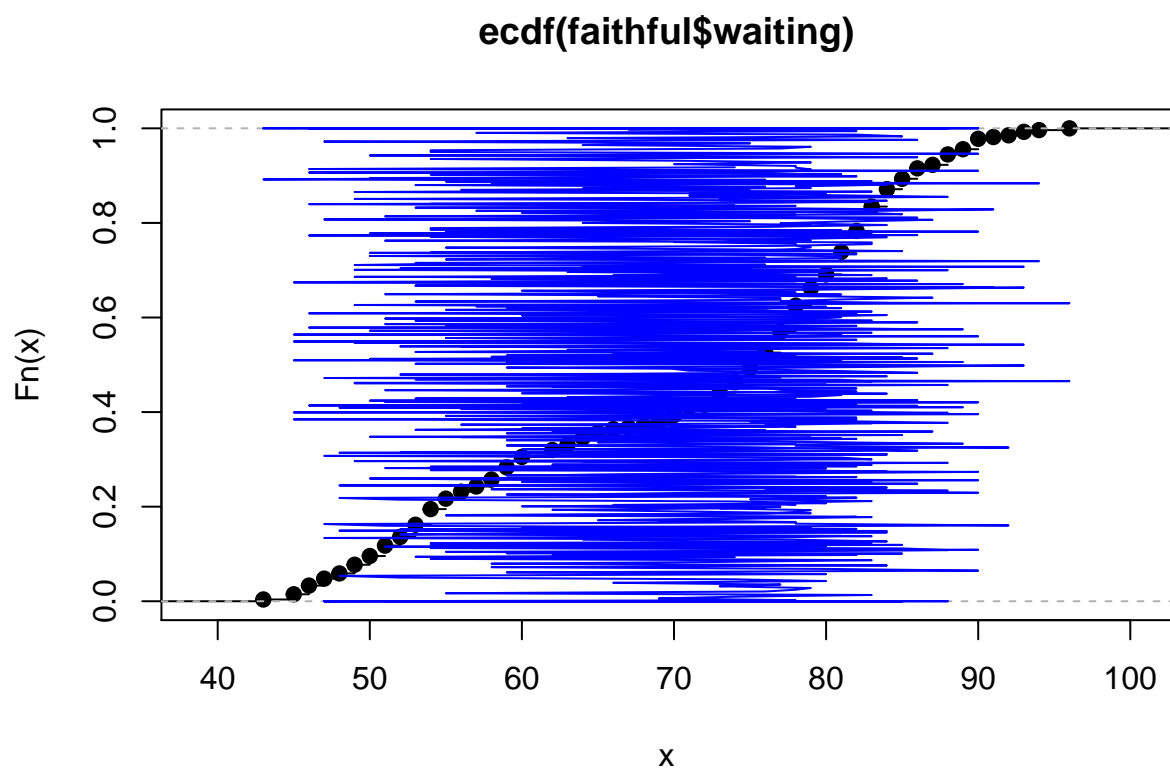


```
plot(ecdf(fijiquakes$mag))
```



```
#lower and upper bands:
n <- sum(!is.na(faithful$waiting))
L<-1:n
U<-1:n
epsilon_i = sqrt(log(2/0.05)/(2*n))

L=pmax(1:n/n-epsilon_i, 0)
U=pmin(1:n/n+epsilon_i, 1)
plot(ecdf(faithful$waiting))
lines(faithful$waiting, U, col="blue")
lines(faithful$waiting, L, col="blue")
```

```
# Calculate  $F(4.9) - F(4.3)$   
ecdf(fijiquakes$mag)(4.9) - ecdf(fijiquakes$mag)(4.3)
```

```
## [1] 0.526
```

```
# Mean and median waiting time  
mean(faithful$waiting)
```

```
## [1] 70.89706
```

```
median(faithful$waiting)
```

```
## [1] 76
```