hw_emperical

LonghaoChen

3/4/2019

1. Is the data in the file maybe uniform.txt distributed as a Uniform distribution on [0, 1]?

```
library(reshape2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- ti
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------ tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fitdistrplus)
```

```
## Warning: package 'fitdistrplus' was built under R version 3.5.2
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: survival
```

```
## Loading required package: npsurv
```

```
## Loading required package: lsei
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

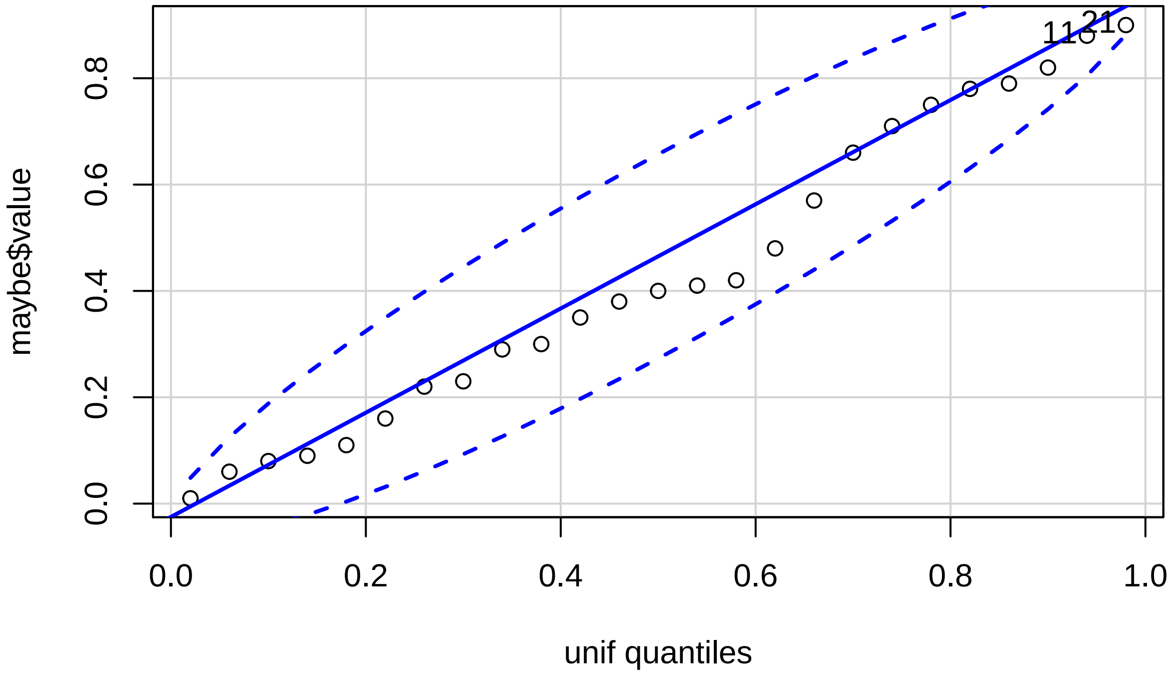
```
maybe_uniform <- read.table("~/Desktop/MA677/HW/maybe_uniform.txt", quote="\"", comment.char="")
```

```
## Warning in read.table("~/Desktop/MA677/HW/maybe_uniform.txt", quote =
## "\"", : incomplete final line found by readTableHeader on '~/Desktop/MA677/
## HW/maybe_uniform.txt'
```

```
maybe <- reshape2::melt(maybe_uniform)
```

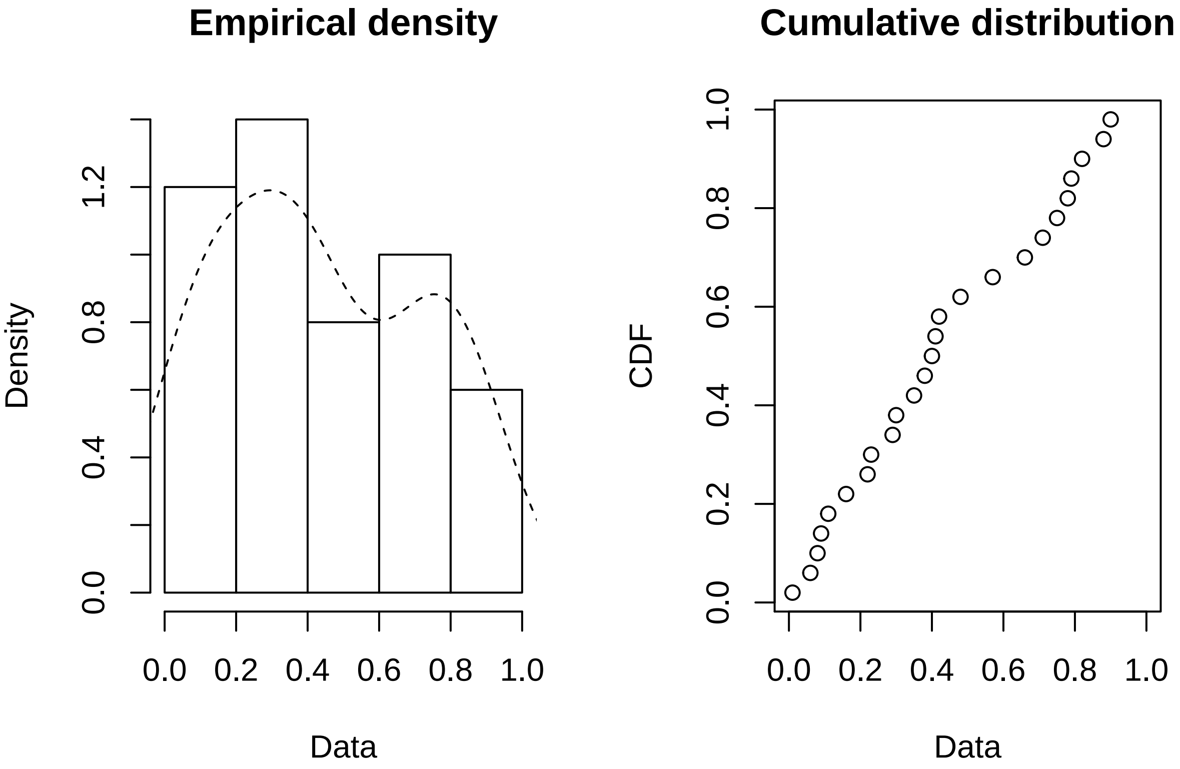
```
## No id variables; using all as measure variables
```

```
qqPlot(maybe$value,distribution = "unif")
```

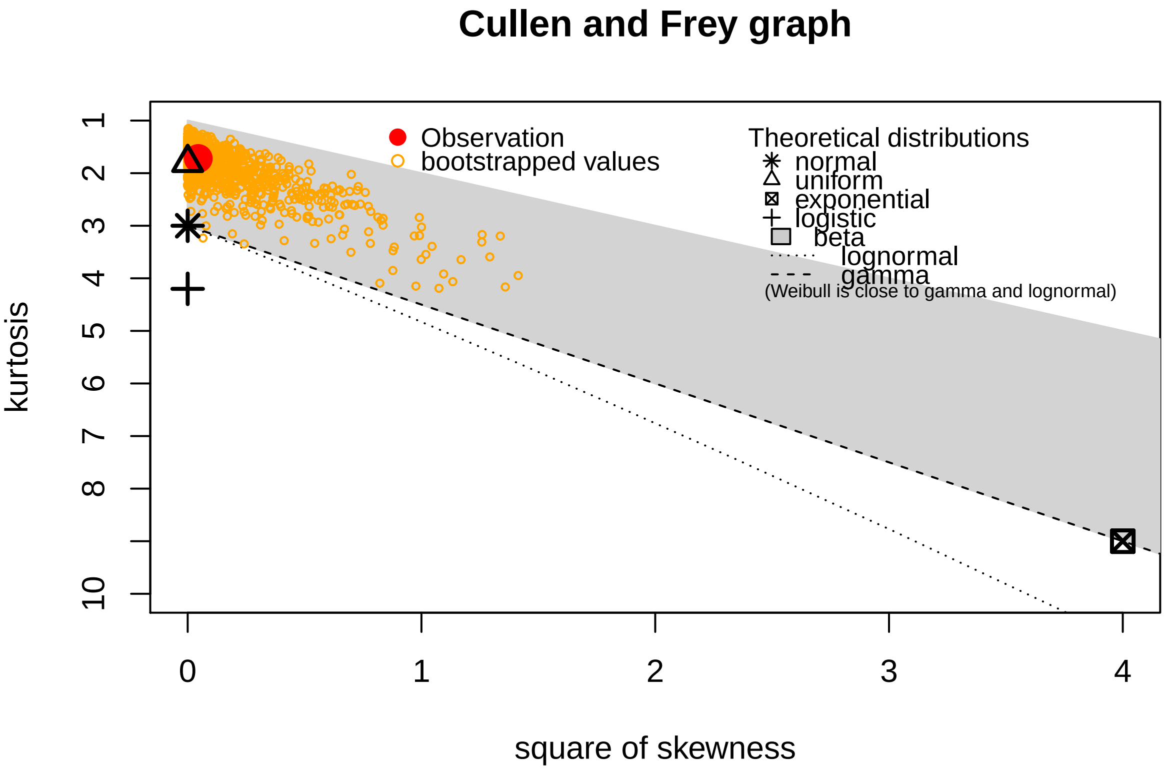


```
## [1] 21 11
```

```
plotdist(maybe$value, demp = TRUE)
```



```
descdist(maybe$value, obs.col = "red",boot=1000)
```



```
## summary statistics
## ------
## min:  0.01   max:  0.9
## median:  0.4
## mean:  0.434
## estimated sd:  0.284356
## estimated skewness:  0.2127721
## estimated kurtosis:  1.721164
```

Yes, I believe it is close to uniform distribution.

Is it possible that the model below is better than the Uniform?

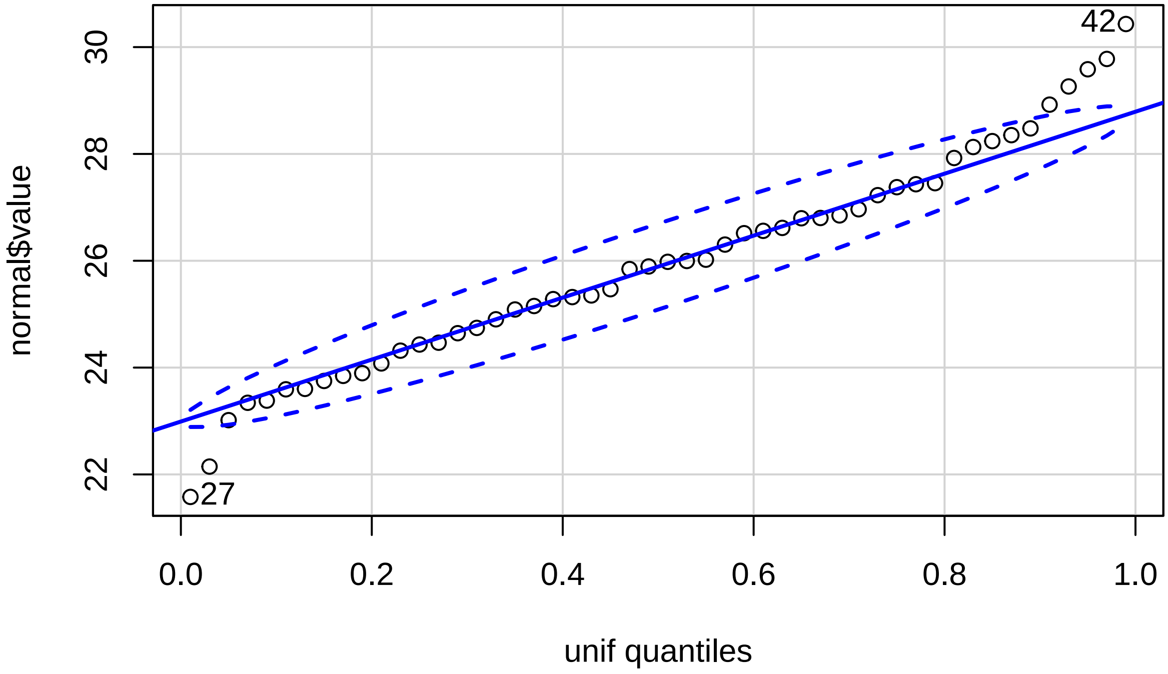
Is there a third model that is a better fit? Not really from our plot.

2. Is the data in the file maybe normal.txt a random sample from the normal distribution with mean = 26 and variance = 4? Investigate your result. Make a qnorm plot. Make a histogram. Be ready to show and discuss your results.

```
maybe_normal <- read.table("~/Desktop/MA677/HW/maybe_normal.txt", quote="\"", comment.char="")
normal <- reshape2::melt(maybe_normal)
```

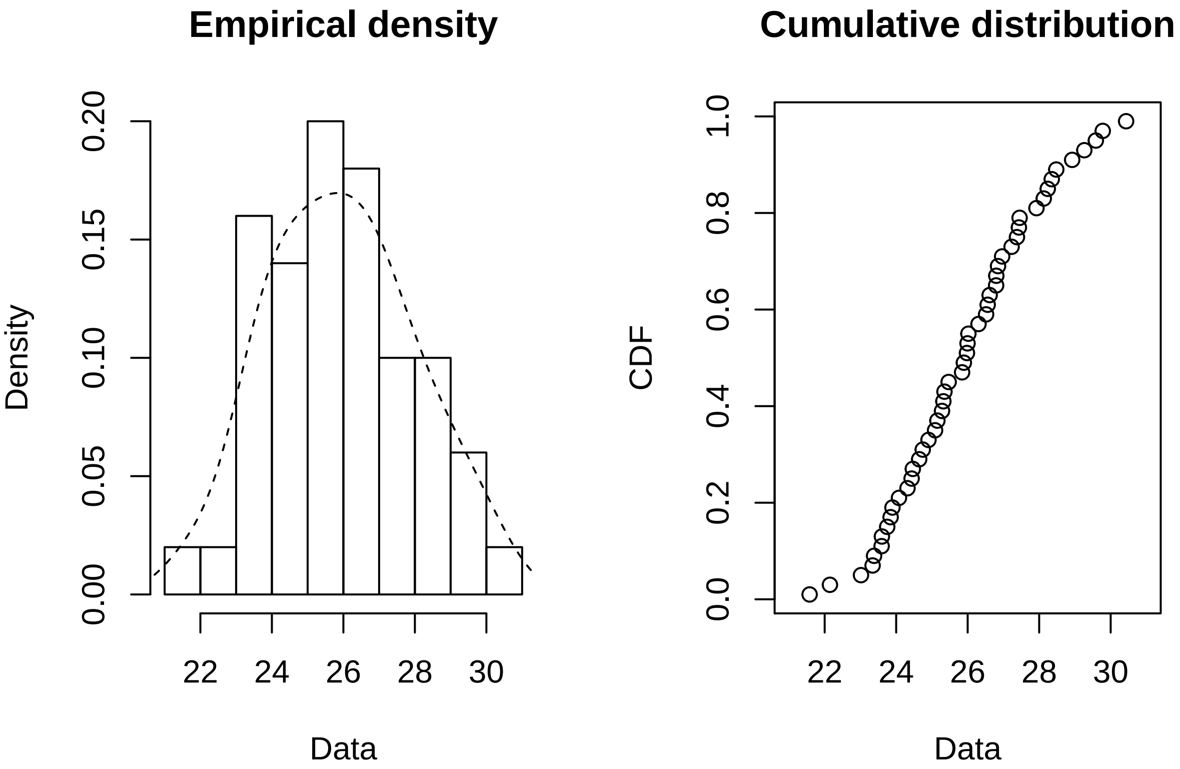
```
## No id variables; using all as measure variables
```

```
qqPlot(normal$value,distribution = "unif")
```

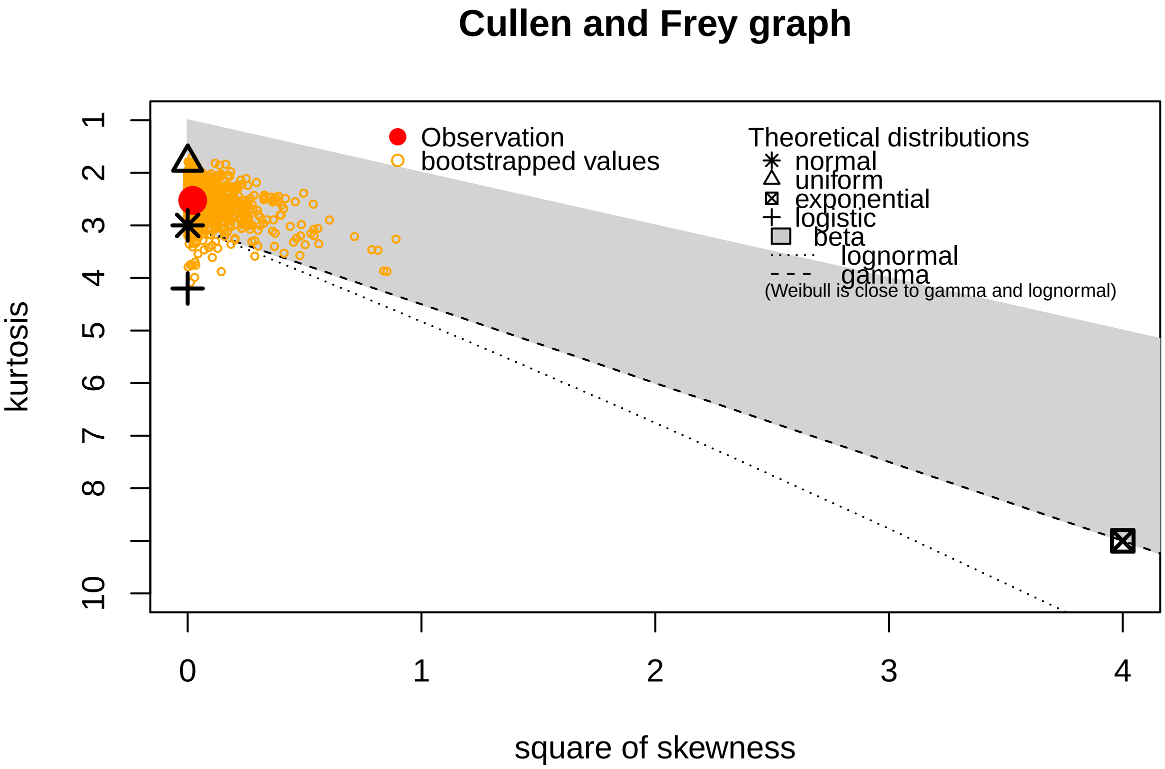


```
## [1] 42 27
```

```
plotdist(normal$value, demp = TRUE)
```



```
descdist(normal$value, obs.col = "red",boot=1000)
```



```
## summary statistics
## ------
## min:  21.579   max:  30.432
## median:  25.9365
## mean:  25.94258
## estimated sd:  2.042374
## estimated skewness:  0.1467526
## estimated kurtosis:  2.525013
```

```
fit <- fitdistr(normal$value, densfun="normal")
fit
```

```
##        mean          sd
##   25.9425800    2.0218472
##  ( 0.2859324) ( 0.2021847)
```

I think it is close to a normal distribution with mean at 26 and variance at 4, since I have evidence from qqplot, histrogram and from fitdistr function.

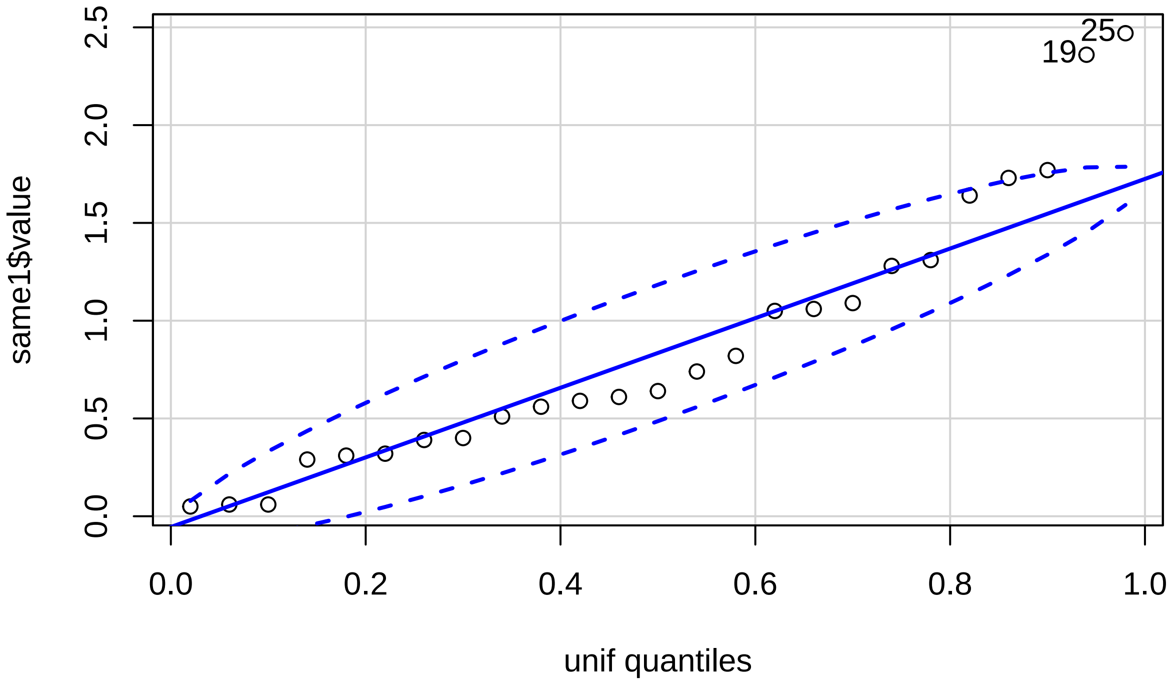
3. Are the two samples in X, maybe same 1.txt, and Y , maybe same 2.txt, from the same distribution? Could it be that X + 2 and Y have the same distribution?

```
maybe_same1 <- read_table2("~/Desktop/MA677/HW/maybe_same_1.txt",
    col_names = FALSE, col_types = cols(X1 = col_number(),
        X2 = col_number(), X3 = col_number(),
        X4 = col_number(), X5 = col_number()))
maybe_same2 <- read_table2("~/Desktop/MA677/HW/maybe_same_2.txt",
    col_names = FALSE, col_types = cols(X1 = col_number(),
        X2 = col_number(), X3 = col_number(),
        X4 = col_number(), X5 = col_number()))
```

```
## Warning: The following named parsers don't match the column names: X5
```

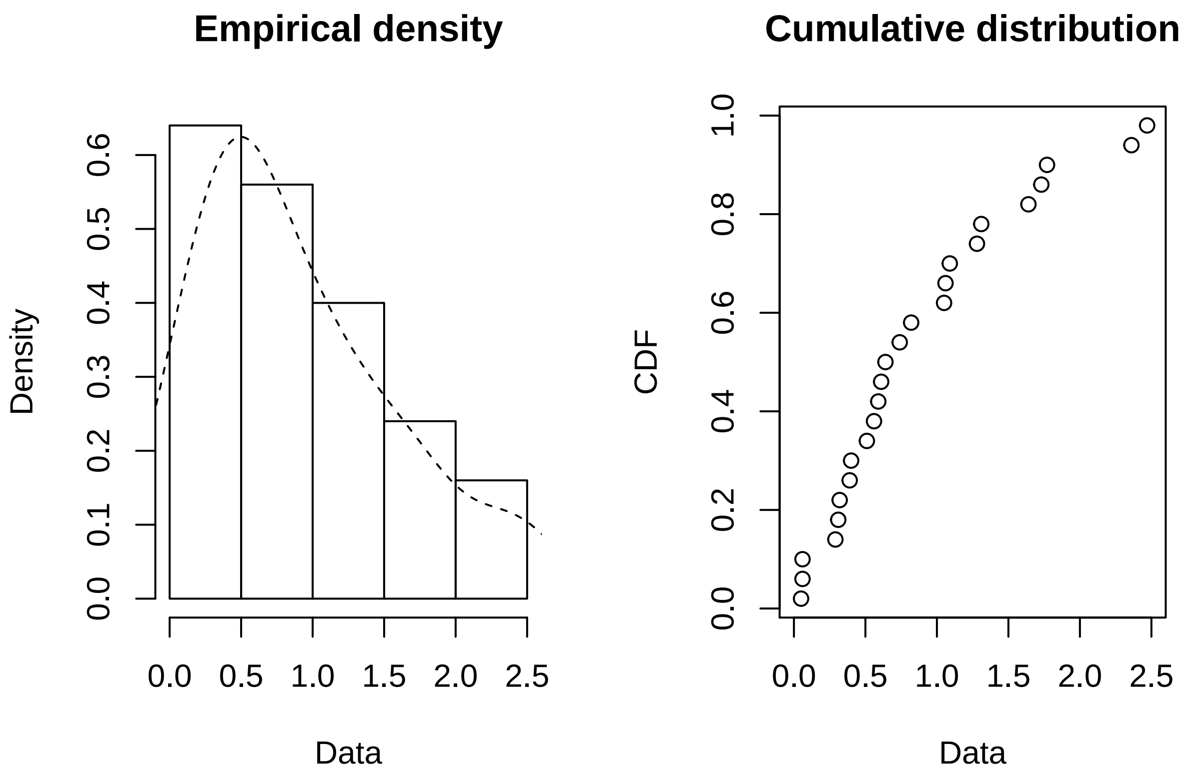
```
same1 <- reshape2::melt(maybe_same1,id.vars=NULL)
same2 <- reshape2::melt(maybe_same2,id.vars=NULL)

qqPlot(same1$value,distribution = "unif")
```

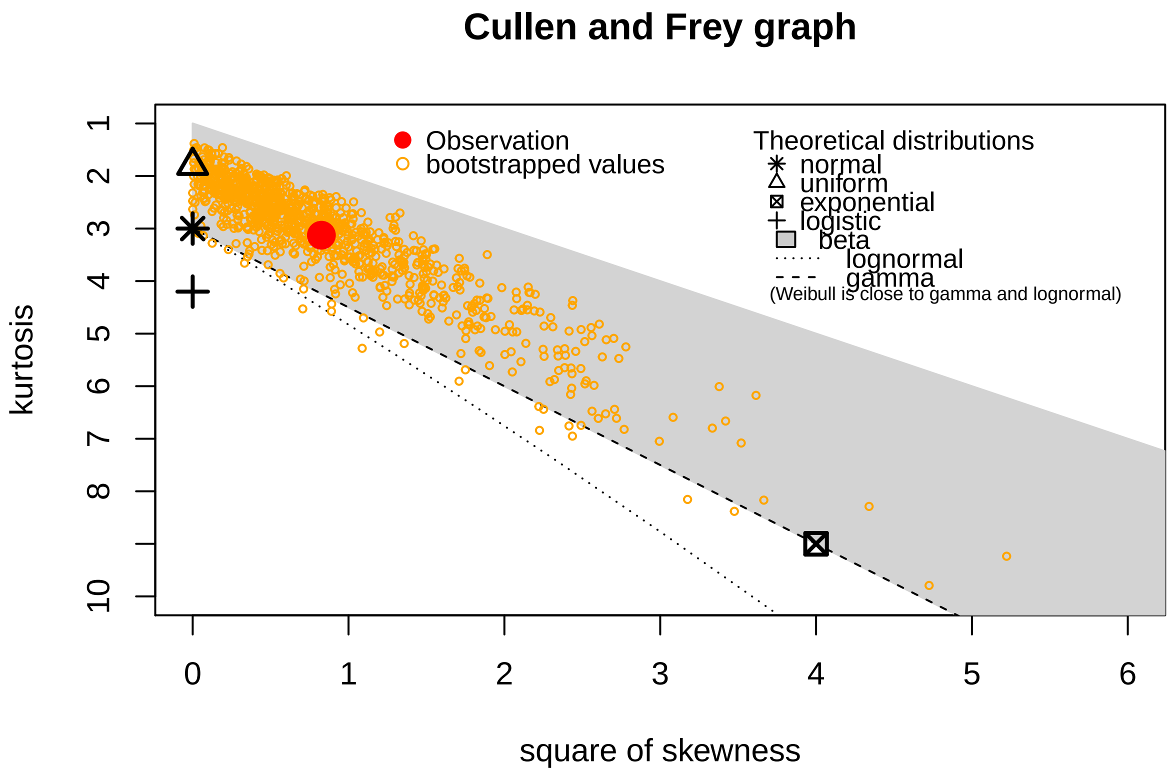


```
## [1] 25 19
```

```
plotdist(same1$value, demp = TRUE)
```

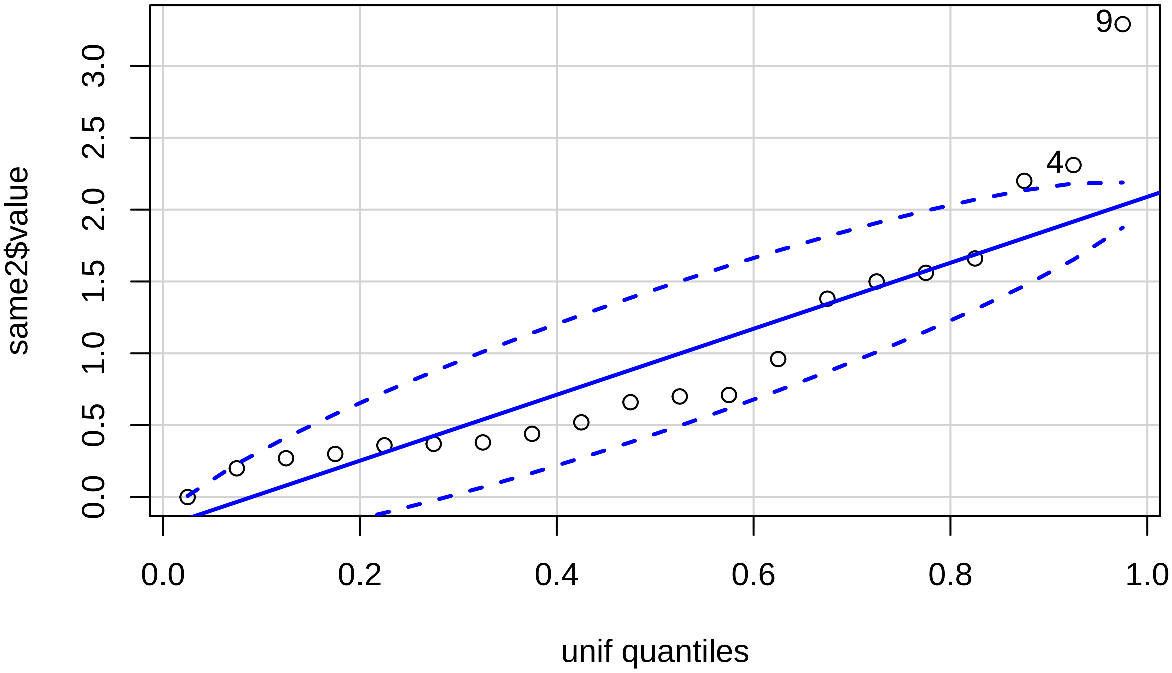


```
descdist(same1$value, obs.col = "red",boot=1000)
```



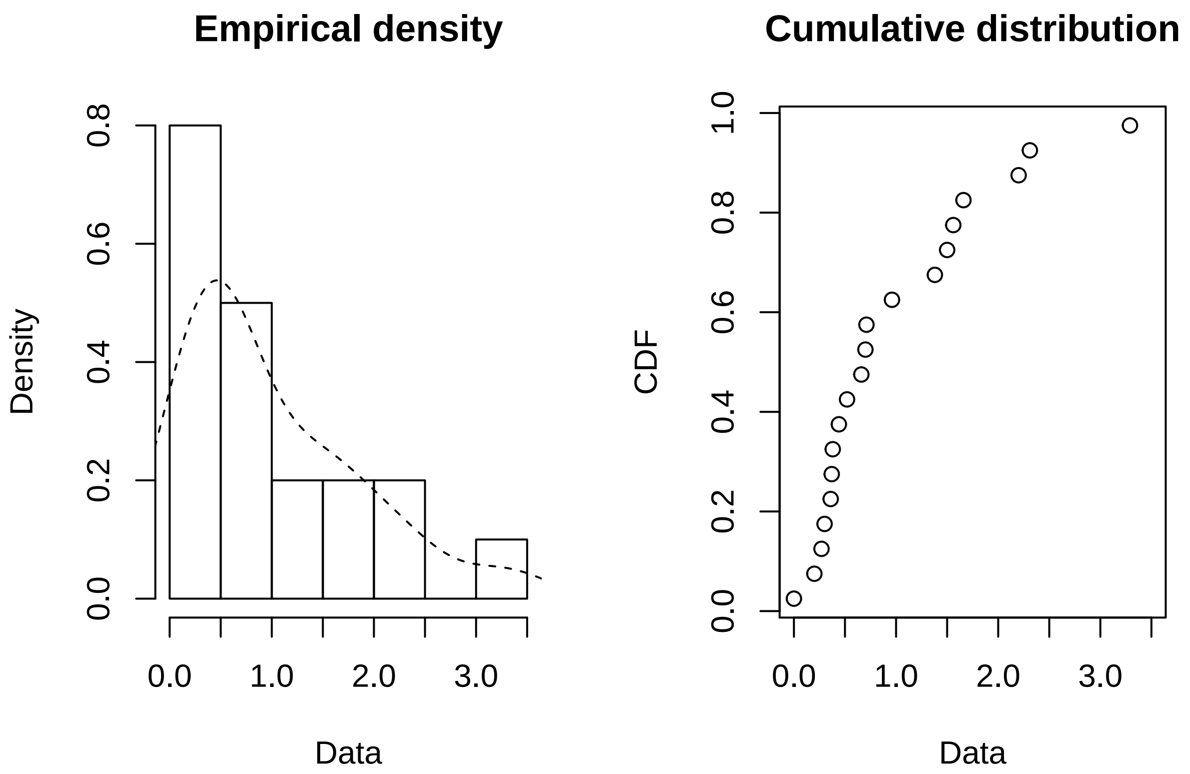
```
## summary statistics
## ------
## min:  0.05   max:  2.47
## median:  0.64
## mean:  0.8844
## estimated sd:  0.6839961
## estimated skewness:  0.9088325
## estimated kurtosis:  3.12617
```

```
qqPlot(same2$value,distribution = "unif")
```

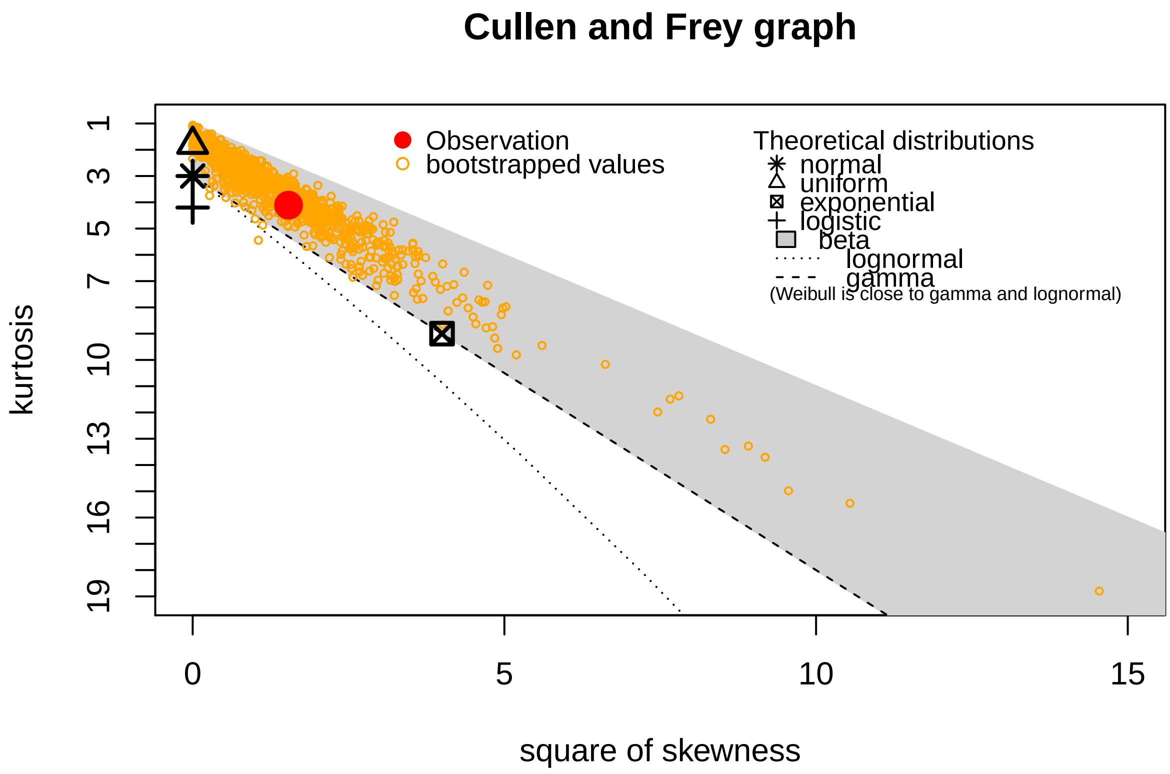


```
## [1] 9 4
```

```
plotdist(same2$value, demp = TRUE)
```

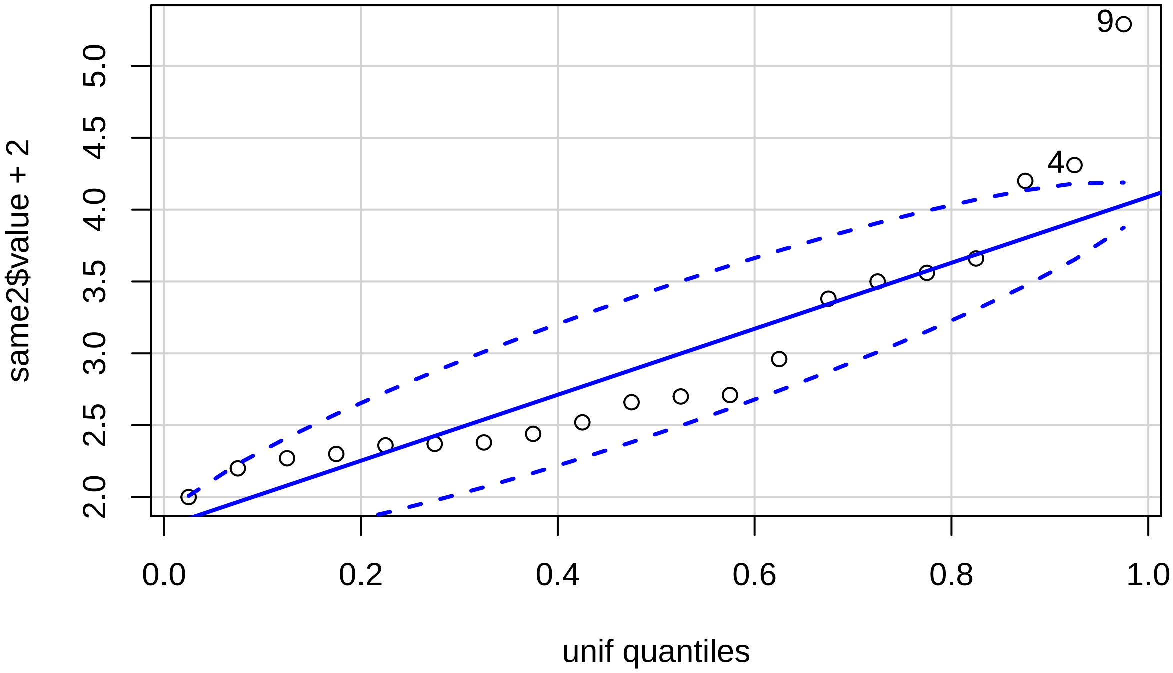


```
descdist(same2$value, obs.col = "red",boot=1000)
```



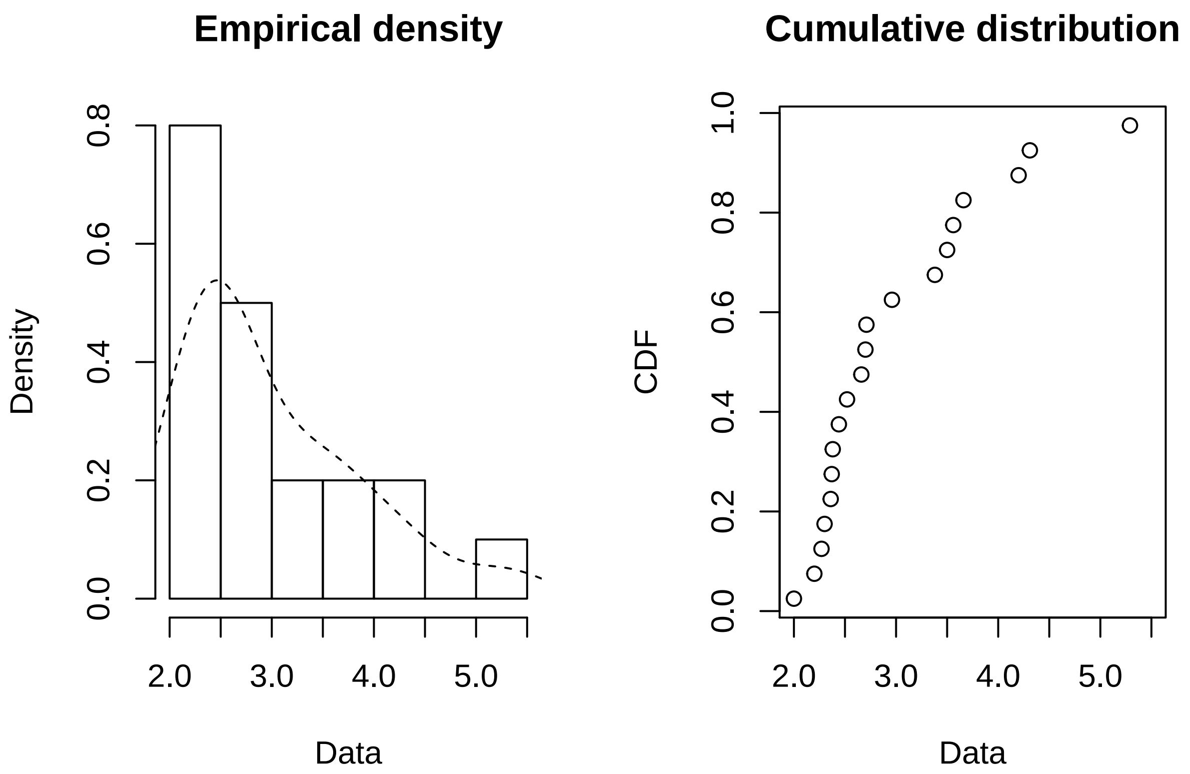
```
## summary statistics
## ------
## min:  0   max:  3.29
## median:  0.68
## mean:  0.9885
## estimated sd:  0.8654252
## estimated skewness:  1.240338
## estimated kurtosis:  4.109623
```

```
qqPlot(same2$value+2,distribution = "unif")
```

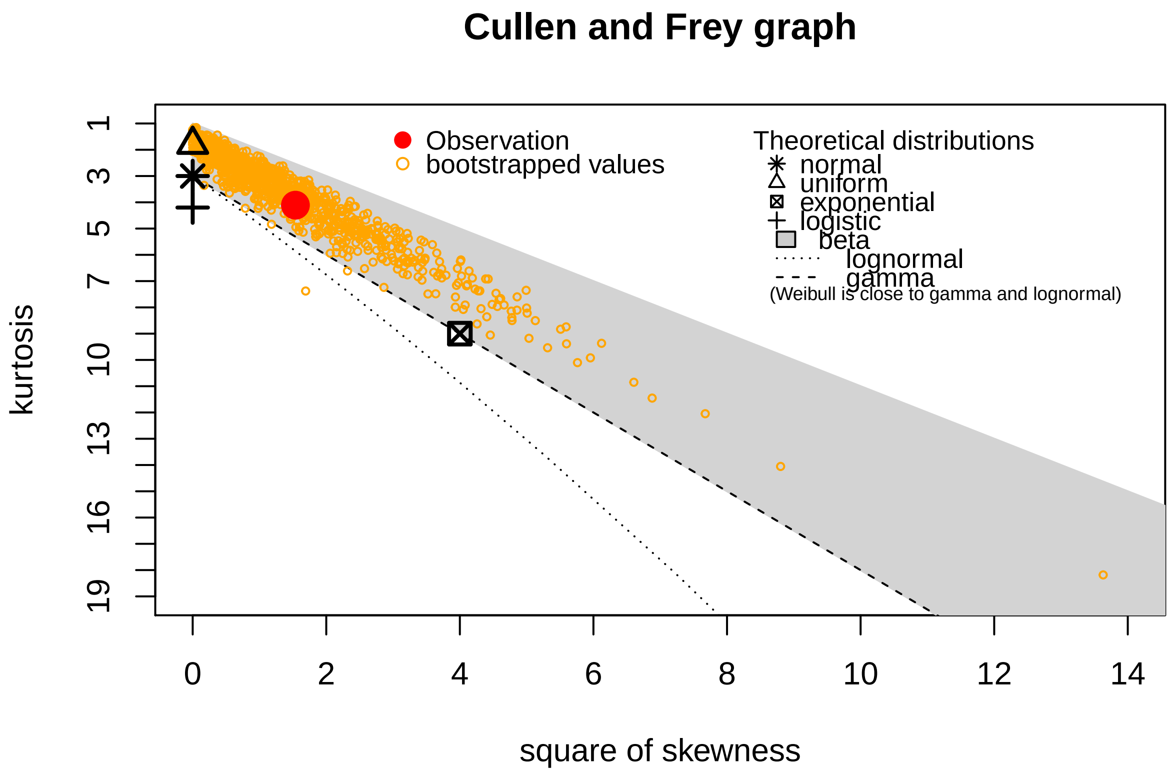


```
## [1] 9 4
```

```
plotdist(same2$value+2, demp = TRUE)
```



```
descdist(same2$value+2, obs.col = "red",boot=1000)
```

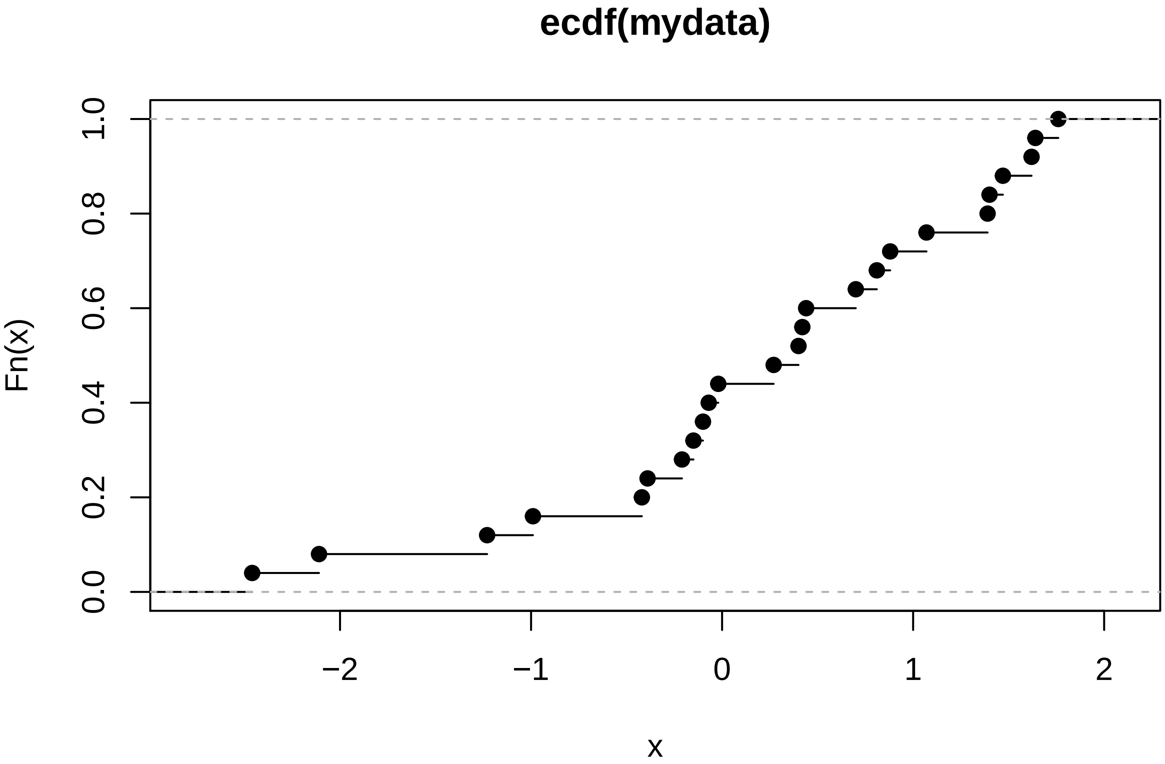


```
## summary statistics
## ------
## min:  2   max:  5.29
## median:  2.68
## mean:  2.9885
## estimated sd:  0.8654252
## estimated skewness:  1.240338
## estimated kurtosis:  4.109623
```

By comparing the plots of X, Y and X+2, I think the the relationship is stronger between X and Y rather than X+2 and Y. Therefore, X and Y are more likely to come from the same distribution.

4. Read the data in the file norm data.Rdata. There are 25 data points. Is this a data set drawn from the **standard normal distribution** Use ecdf() to compute the empirical distribution of the data. Create a normal distribution that can be used to calculate the KolmogorovSmirnov test. Calculate the D statistic. Run the ks.test() function and compare your results to the results reported by ks.test.

```
mydata <- readRDS("~/Desktop/MA677/HW/norm_sample.Rdata")
b <- ecdf(mydata)
plot(b)
```



```
a <- rnorm(n = 25,mean = 0,sd = 1)
ks.test(x = mydata,y = a)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  mydata and a
## D = 0.24, p-value = 0.4755
## alternative hypothesis: two-sided
```

Our results indicate that the data is not standard normal distribution.

5. Produce empirical distributions with confidence bands for the fujiquakes.dat and faithful.dat. For the fujiquakes data, Find a 95for F(4.9) − F(4.3). For the faithful data, estimate a 90 percent confidence interval for the mean waiting time and estimate the median waiting time.