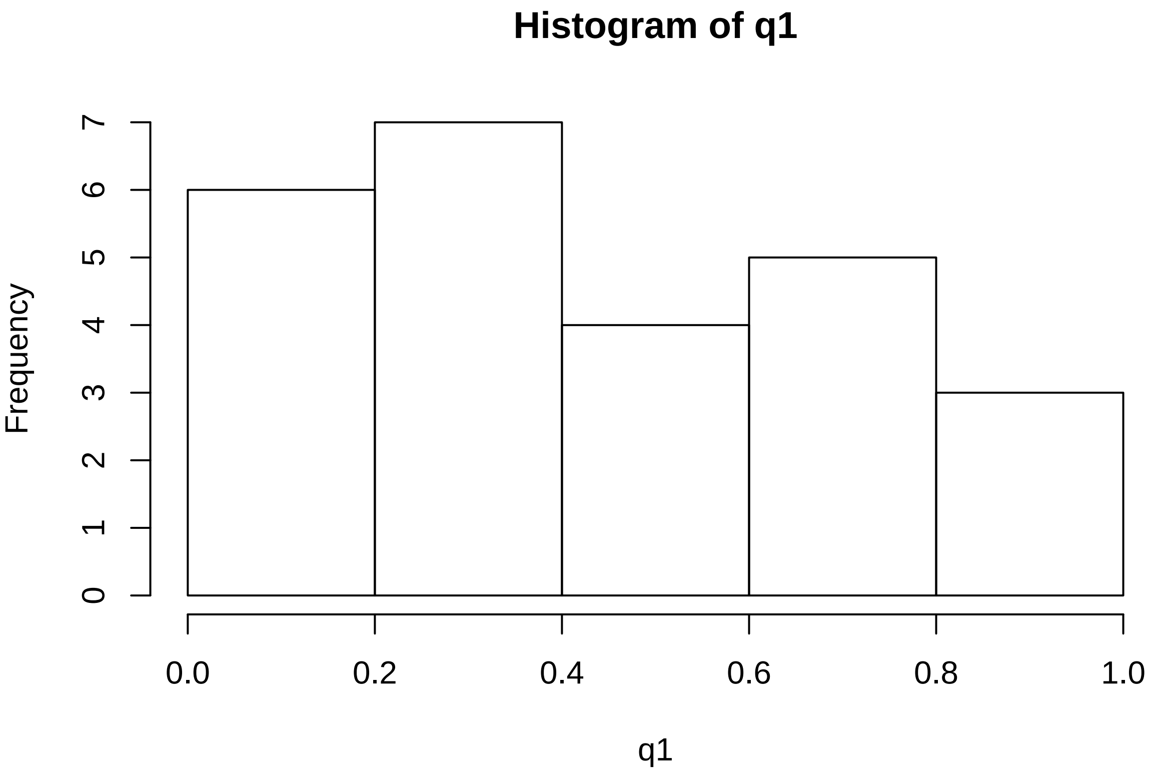
MA677-Assignment3

Stella Li

3/4/2019

1.Maybe Uniform
(a) Check Unif(0,1)

```
q1 <- read.table("maybe_uniform.txt");q1 <- unlist(q1)
hist(q1)
```



```
ks.test(q1, "punif")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  q1
## D = 0.18, p-value = 0.3501
## alternative hypothesis: two-sided
```

Taking a look at the histogram and the KS test result, it seems that the data is not from uniform distribution.

(b) Check new model

```
q1_d <- ecdf(q1)
tr_d <- rep(0, 12)
for(i in 1:25) tr_d[i] = ifelse(q1[i] < .5, 1.5*q1[i], .75+(q1[i]-.5)*.5)
max(abs(q1_d(q1) - tr_d))
```

```
## [1] 0.11
```

The calculated D is 0.11, less than the D from Uniform distribution. Therefore, the new distribution is better.

(c) Try a gamma distribution with rate = 2 and shape = 1.

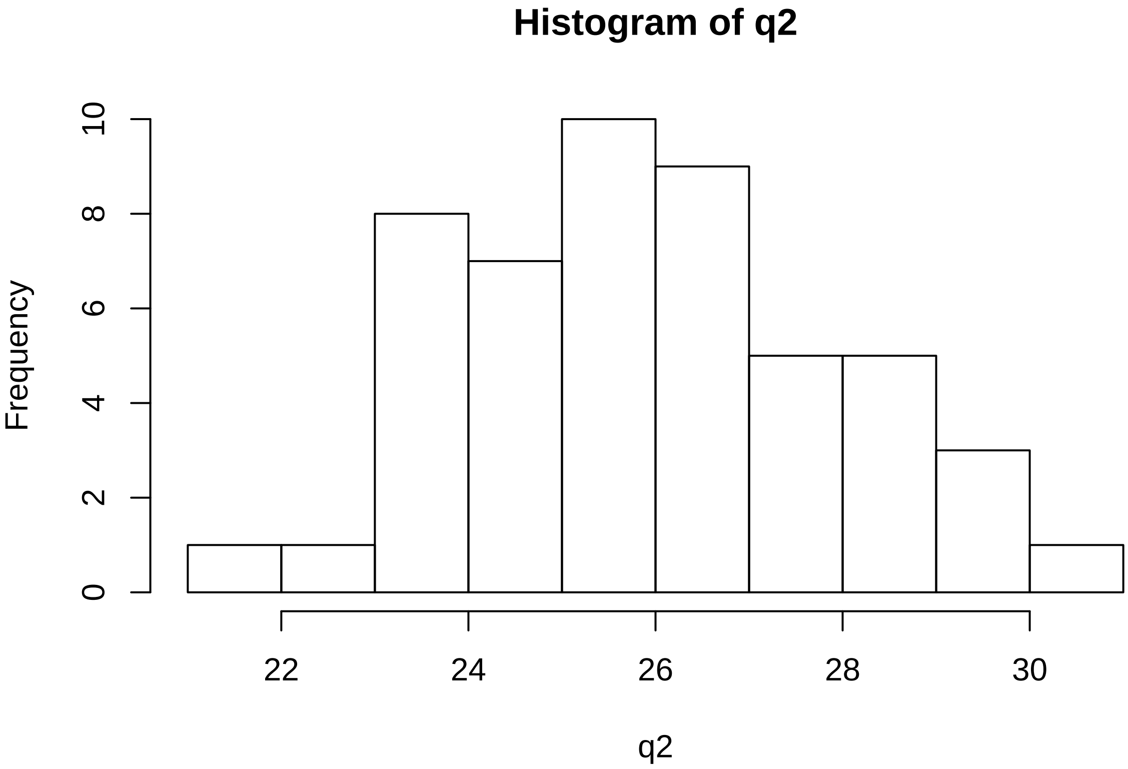
```
ks.test(q1, "pgamma", rate = 2, shape = 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  q1
## D = 0.1653, p-value = 0.4535
## alternative hypothesis: two-sided
```

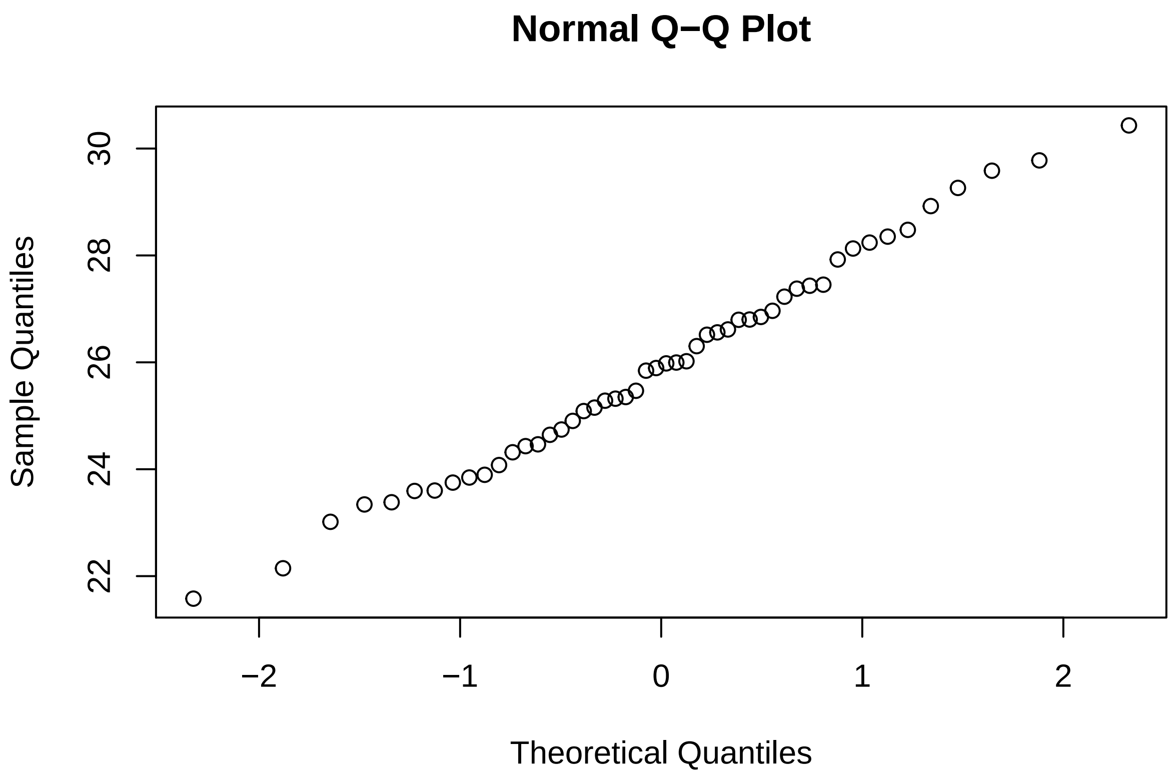
The result shows that this is better than the uniform distribution, but not as good as the model in (b)

2. Maybe normal

```
q2 <- read.table("maybe_normal.txt");q2 <- unlist(q2)
hist(q2)
```



```
qqnorm(q2)
```



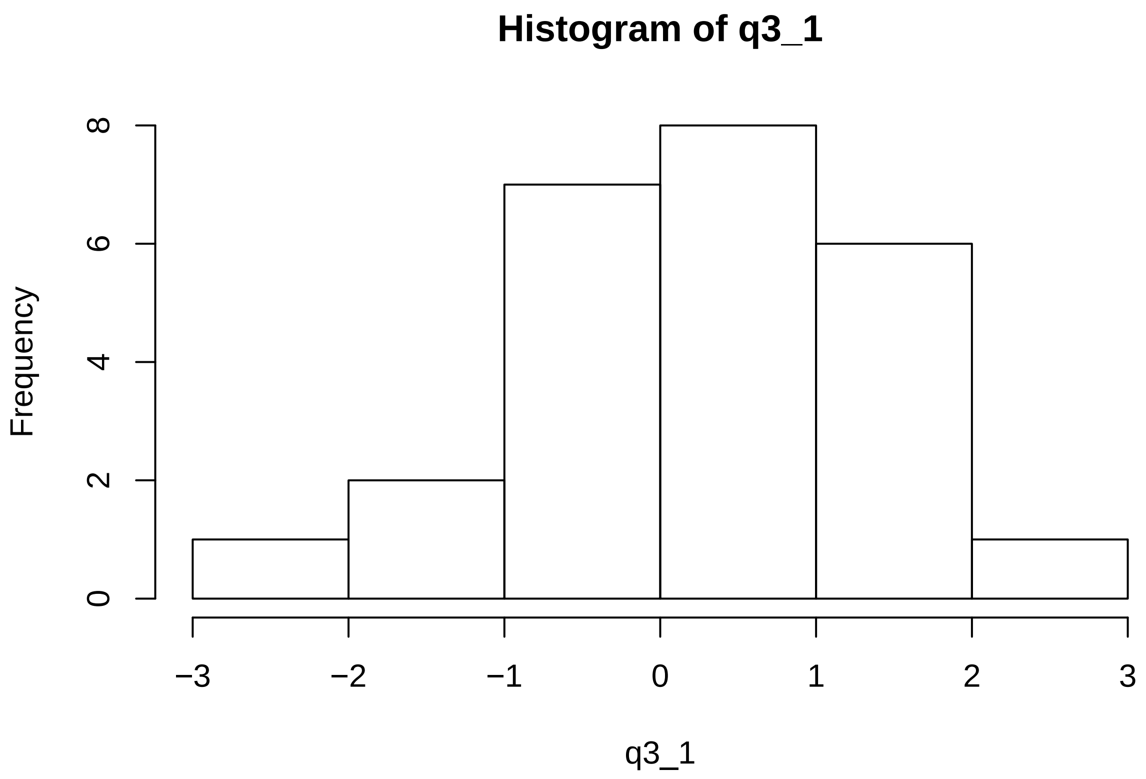
```
ks.test(q2, "pnorm", mean = 26, sd = sqrt(4))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  q2
## D = 0.06722, p-value = 0.9663
## alternative hypothesis: two-sided
```

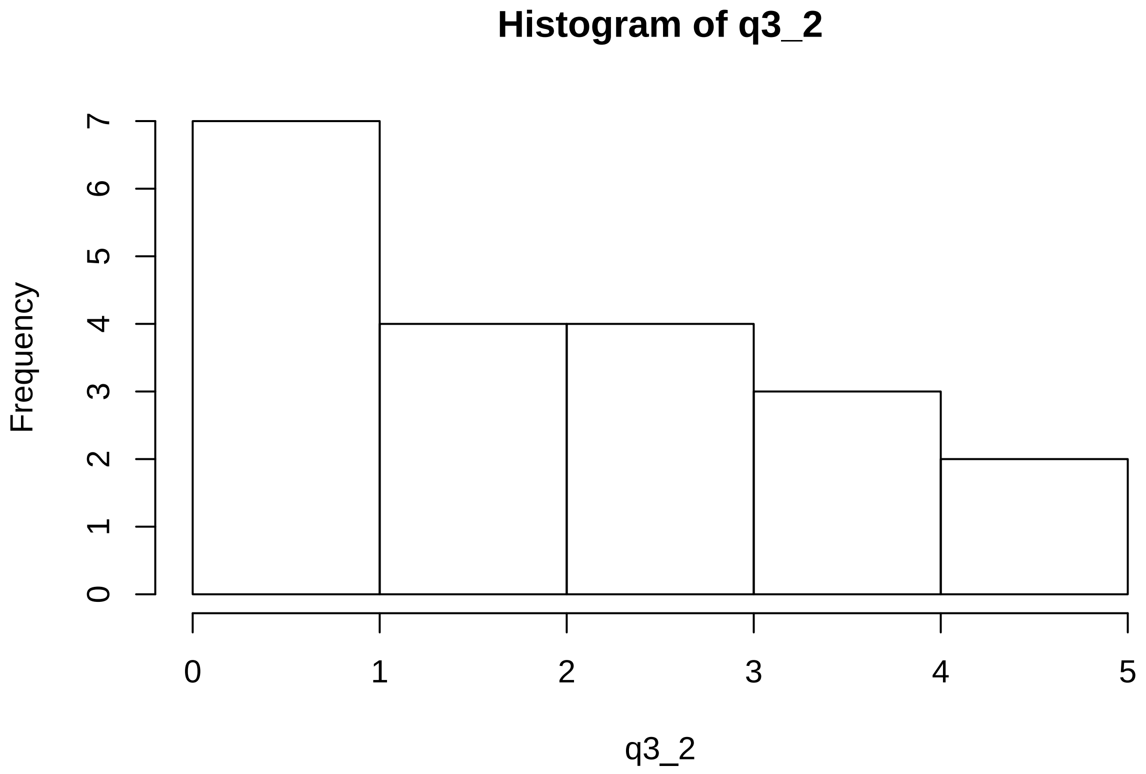
It is very likely that the data is from normal distribution with mean $= 26$ and variance $= 4$.

3. Maybe same

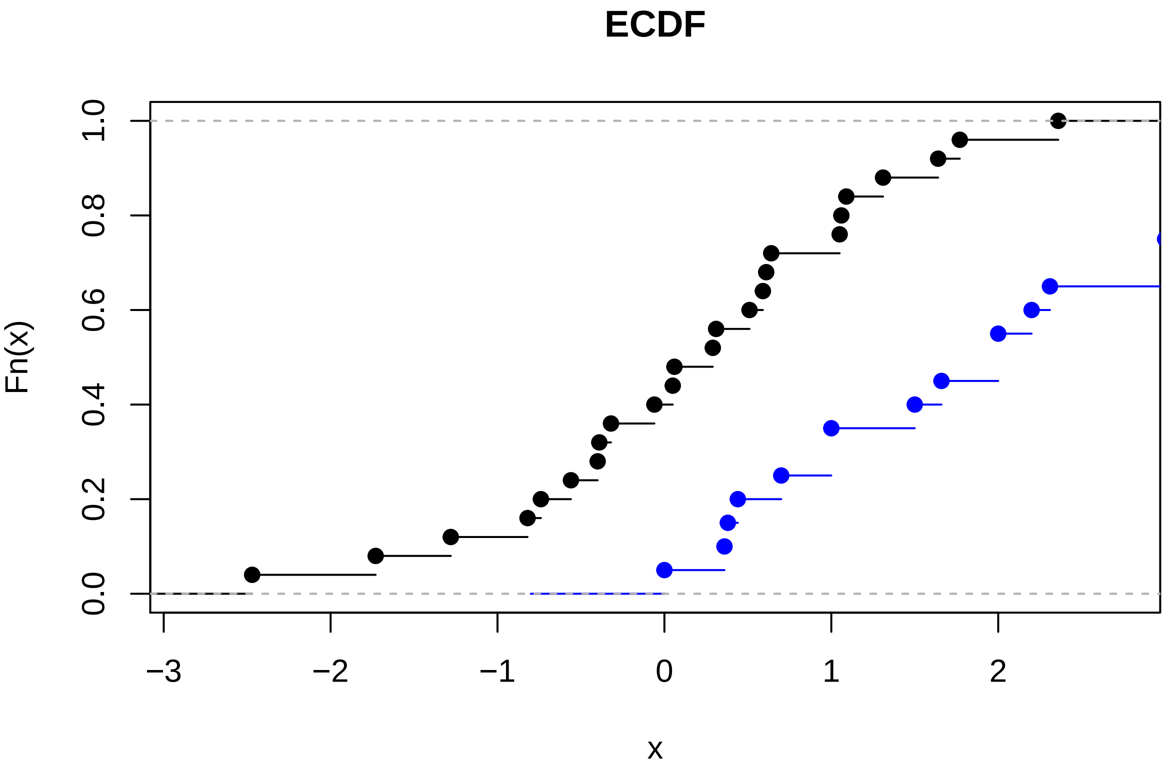
```
q3_1 <- read.table("maybe_same_1.txt"); q3_1 <- unlist(q3_1)
q3_2 <- read.table("maybe_same_2.txt"); q3_2 <- unlist(q3_2)
hist(q3_1)
```



```r
hist(q3_2)
```



```r
plot(ecdf(q3_1), main = "ECDF"); plot(ecdf(q3_2), add = T, col = "blue")
```

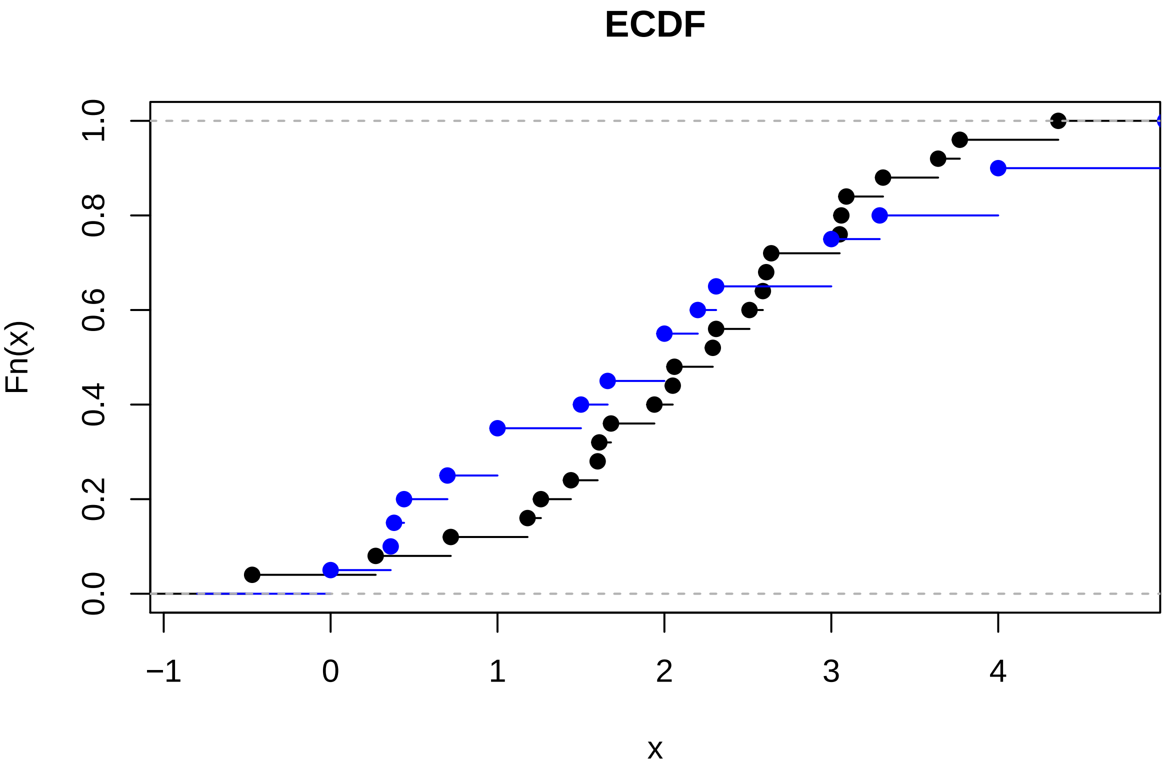


```
ks.test(q3_1, q3_2)
```

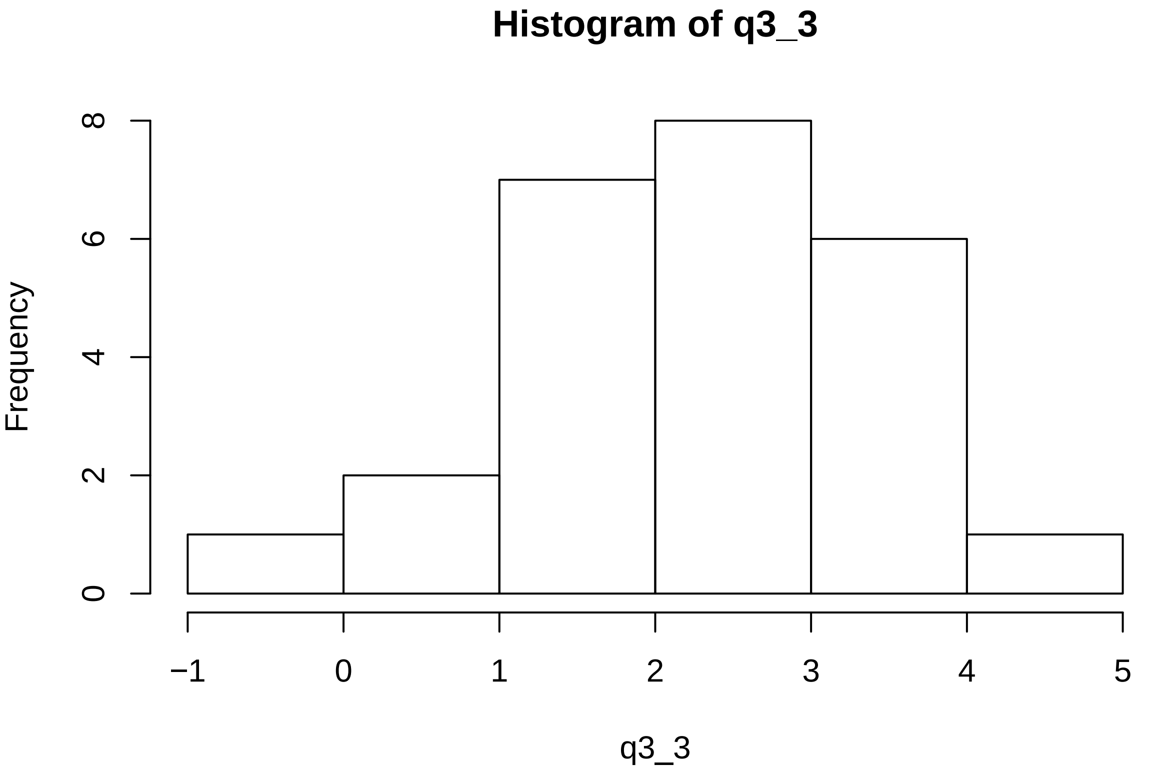
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  q3_1 and q3_2
## D = 0.53, p-value = 0.003891
## alternative hypothesis: two-sided
```

Based on the histograms and test result, these two sets of data do not come from the same distribution.

```
q3_3 <- q3_1 + 2
plot(ecdf(q3_3), main = "ECDF"); plot(ecdf(q3_2), add = T, col = "blue")
```



```r
hist(q3_3)
```



```r
ks.test(q3_3, q3_2)
```

```
## Warning in ks.test(q3_3, q3_2): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  q3_3 and q3_2
## D = 0.23, p-value = 0.5992
## alternative hypothesis: two-sided
```

This time, we cannot reject the null hypothesis that X + 2 and Y do not come from the same distribution. However, by taking a look at the histograms, it seems that X + 2 and Y have very different density distribution.

4. Normal data

```r
q4 <- readRDS("norm_sample.Rdata")
q4 <- as.data.frame(q4); colnames(q4) <- "x"
q4_d <- ecdf(q4$x)
q4$ecdf <- q4_d(q4$x)
q4$nm <- pnorm(q4$x)
q4$D <- q4$ecdf - q4$nm
max(abs(q4$D))
```

```
## [1] 0.1372427
```

```r
ks.test(q4$x, "pnorm", mean = 0, sd = 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  q4$x
## D = 0.17724, p-value = 0.3683
## alternative hypothesis: two-sided
```

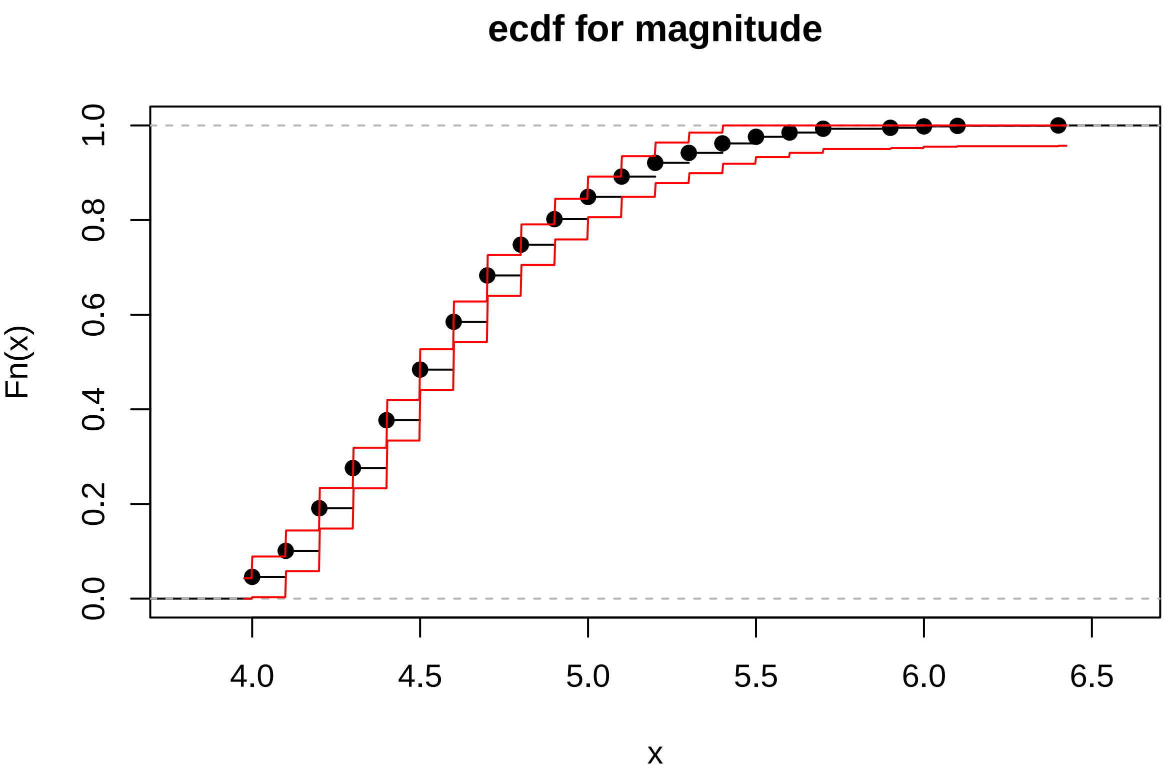
Based on the test result, we fail to reject the hypothesis that the data are drawn from standard normal distribution.
The D statistics from ks.test() function is a little larger than the D we calculated.

5. Fiji and Faithful data

```r
#fiji
fiji <- read.table("fijiquakes.dat", header = T)
fiji_d <- ecdf(fiji$mag)
print(diff_mean <- fiji_d(4.9) - fiji_d(4.3))
```

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

```r
n <- length(fiji$mag)
alpha <- .05
epsl <- sqrt(log(2/alpha)/n/2)
r<-max(fiji$mag) - min(fiji$mag)
grid<-seq(from=min(fiji$mag)-0.01*r,to=max(fiji$mag)+0.01*r,l=1000)
low.cdf<-pmax(fiji_d(grid)-epsl,0)
up.cdf<-pmin(fiji_d(grid)+epsl,1)
plot(fiji_d, main = "ecdf for magnitude");lines(grid,low.cdf,col="red");lines(grid,up.cdf,col="red")
```

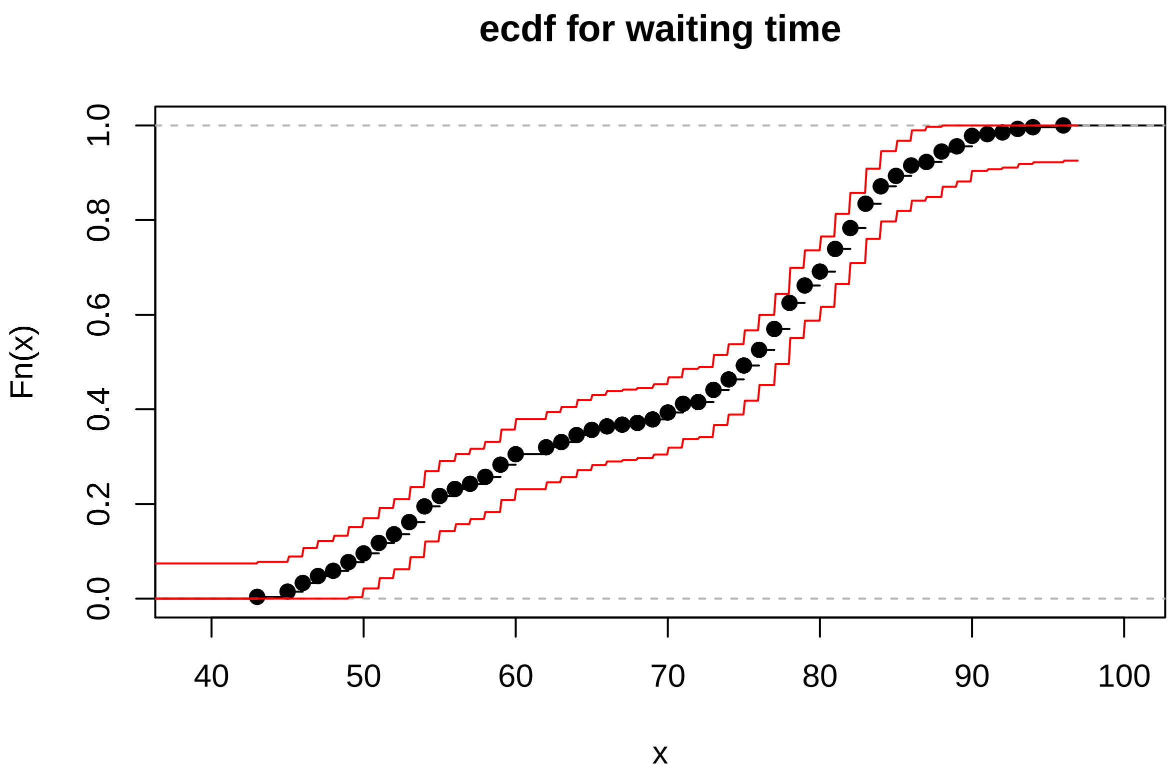


```
### 95% confidence interval for F(4.9) - F(4.3)

tot<-sum((fiji$mag<=4.9) & (fiji$mag>4.3))
binconf(tot,length(fiji$mag),method="wilson",alpha)
```

```
##  PointEst     Lower     Upper
##     0.526 0.4950118 0.5567892
```

```
#--------------------------------------------------------------------
#faithful
faithful <- read.table("faithful.dat", header = T, skip = 25)
f_d <- ecdf(faithful$waiting)
n <- length(faithful$waiting)
alpha <- .1
epsl <- sqrt(log(2/alpha)/n/2)
r<-max(faithful)-min(faithful)
grid<-seq(from=min(faithful)-0.01*r,to=max(faithful)+0.01*r,l=1000)
low.cdf<-pmax(f_d(grid)-epsl,0)
up.cdf<-pmin(f_d(grid)+epsl,1)
plot(f_d, main = "ecdf for waiting time");lines(grid,low.cdf,col="red");lines(grid,up.cdf,col="red")
```



```
### 90% confidence interval for mean
mn <- mean(faithful$waiting)
se <- sd(faithful$waiting)/sqrt(n)
print(paste0("90% confidence interval for mean waiting time is [",
             round(mn-1.64*se,2),", ", round(mn+1.64*se,2),"]"))
```

```
## [1] "90% confidence interval for mean waiting time is [69.55, 72.25]"
```

```
### median waiting time
summary(f_d)
```

```
## Empirical CDF:     51 unique values with summary
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   43.00   56.50   70.00   69.67   82.50   96.00
```

Median waiting time is 70.