Homework 3

Frank

3/8/2019

Question1

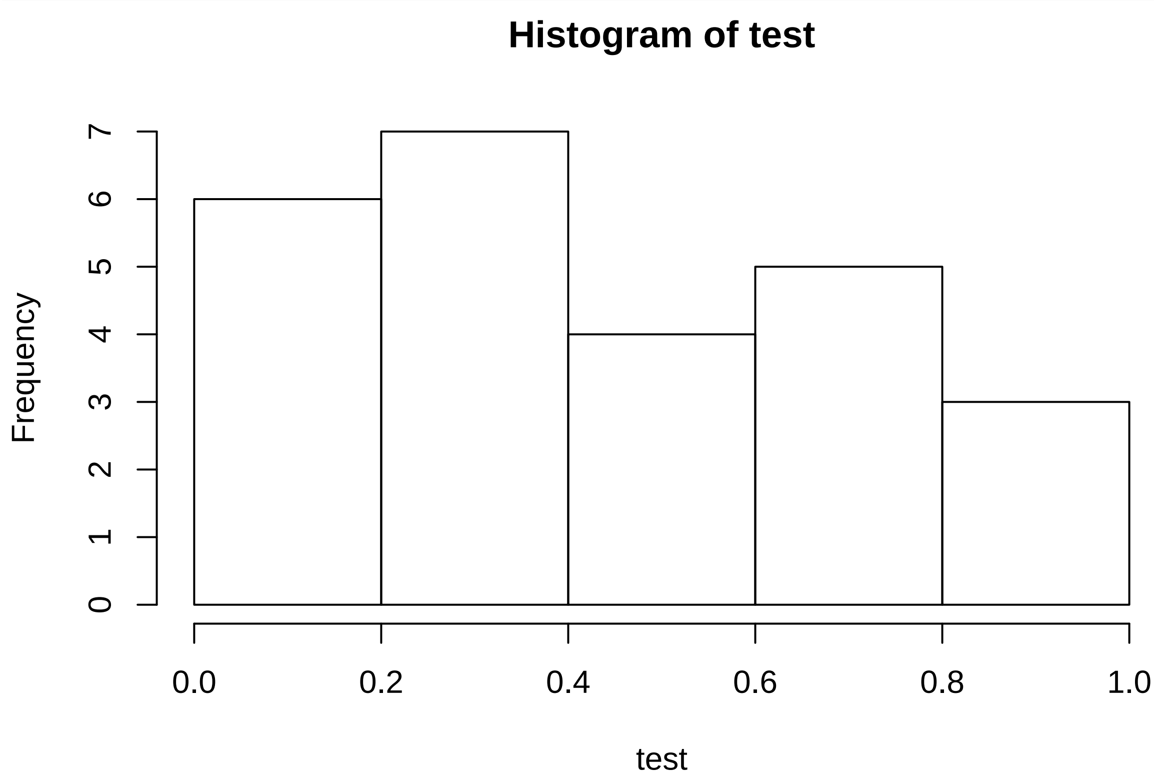
Is the data in the file maybe_uniform.txt distributed as a Uniform distribution on [0, 1]? Is it possible that the model below is better than the Uniform?

Is there a third model that is a better fit?

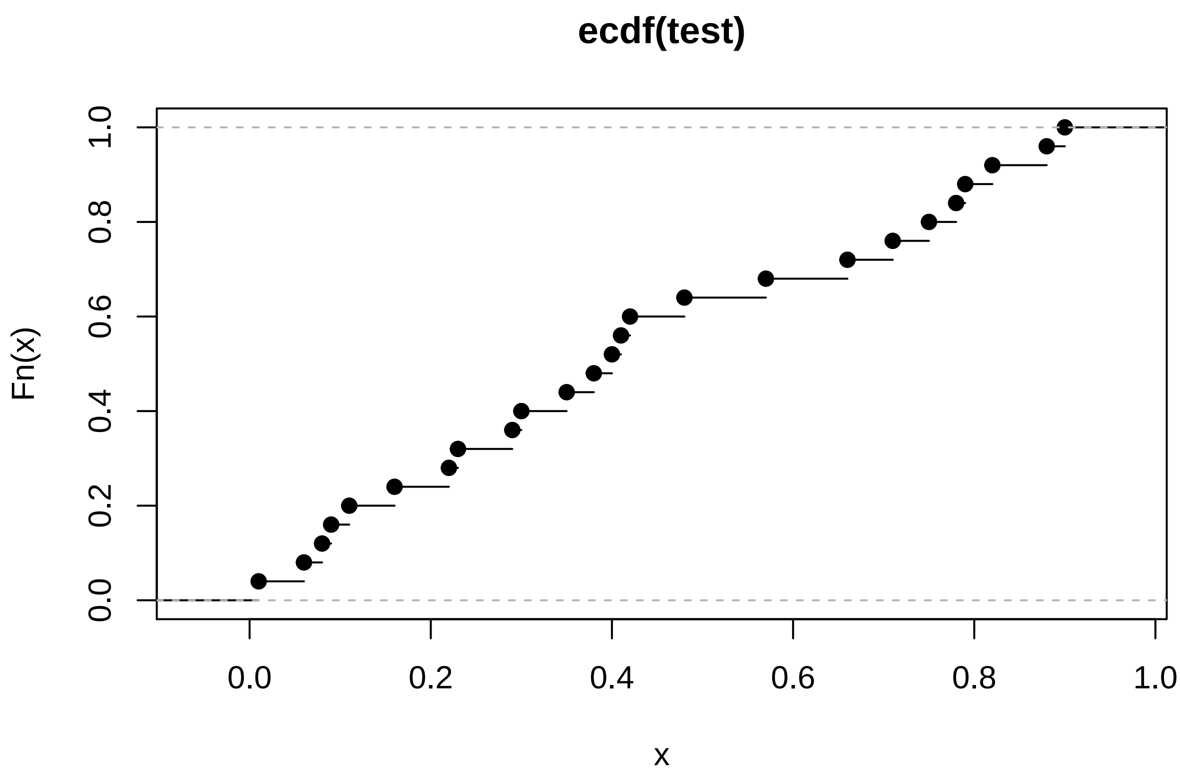
```
maybe_unifrom=read.table("maybe_uniform.txt")
```

```
## Warning in read.table("maybe_uniform.txt"): incomplete final line found by
## readTableHeader on 'maybe_uniform.txt'
```

```
test=c(maybe_unifrom$V1,maybe_unifrom$V2,maybe_unifrom$V3,maybe_unifrom$V4,maybe_unifrom$V5)

hist(test)
```



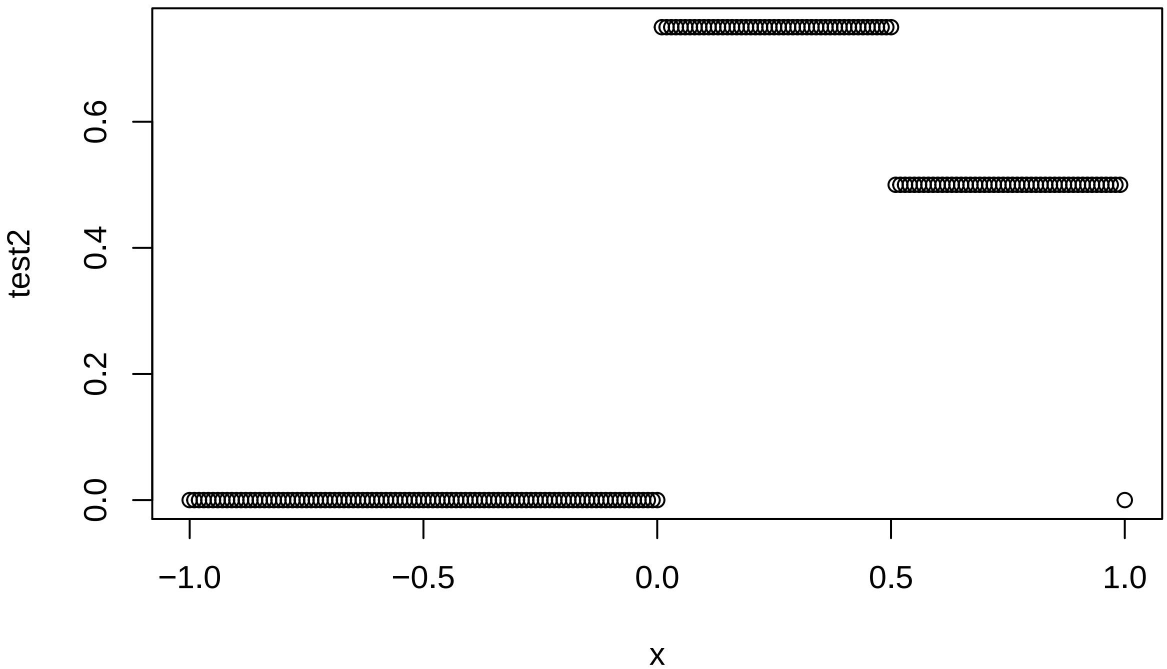

```
#empirical distribution
plot1 <- ecdf(test)
plot(plot1)
```




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

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

```r
x<-seq(-1, 1, by=0.01)

test2 <- ifelse(x > 0 & x <=0.5, 3/4,
   ifelse(x > 0.5 & x < 1,  0.5, 0))
plot(x,test2)
```




```r
ks.test(test2,"punif")
```

```
## Warning in ks.test(test2, "punif"): ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  test2
## D = 0.50746, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

As we can see from the ecdf plot as well as the kstest, the data is more likely to distributed as a Uniform distribution on [0,1].

Question 2

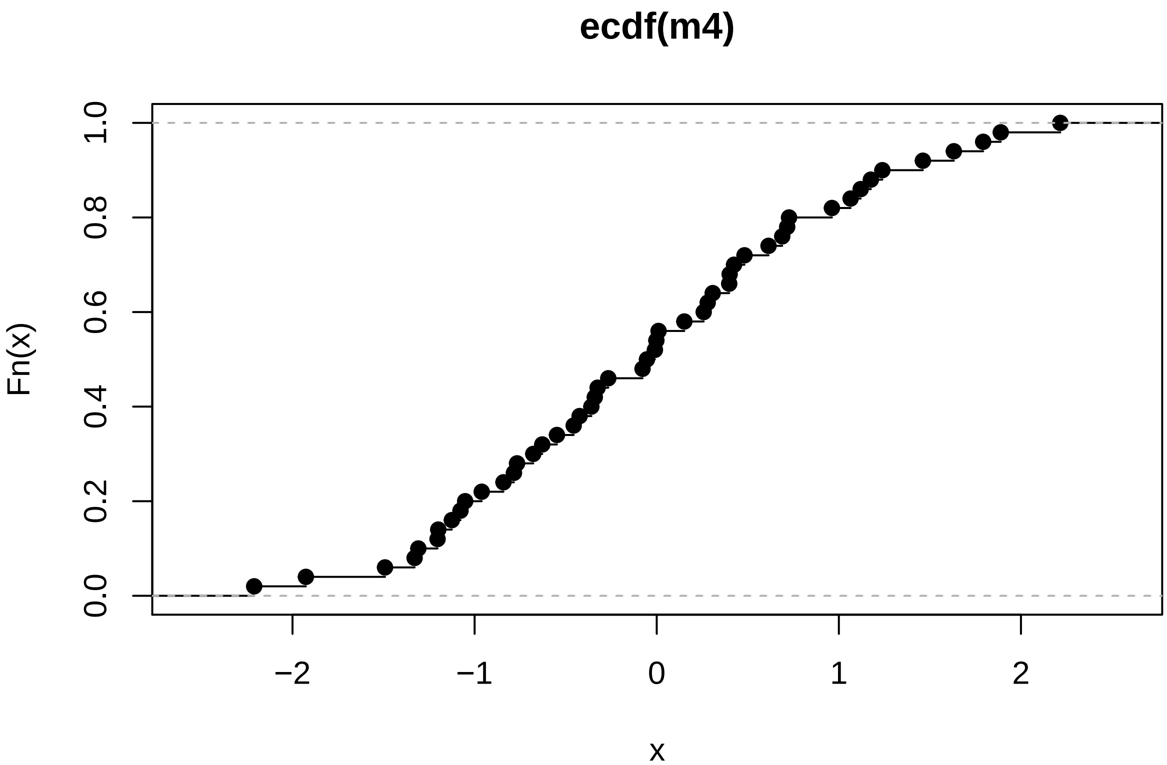
Is the data in the file maybe_normal.txt a random sample from the normal distribution with mean = 26 and variance = 4? Investigate your result. Make a qnorm plot. Make a histogram. Be ready to show and discuss your results.

```r
maybe_normal <- read.table("maybe_normal.txt")
maybe_normal2 <- c(maybe_normal$V1,maybe_normal$V2,maybe_normal$V3,maybe_normal$V4,maybe_normal$V5)

maybe_normal3 <- data.frame(maybe_normal2)

m3 <- maybe_normal2-26
m4 <- m3/2

ks.test(m4,"pnorm")
```

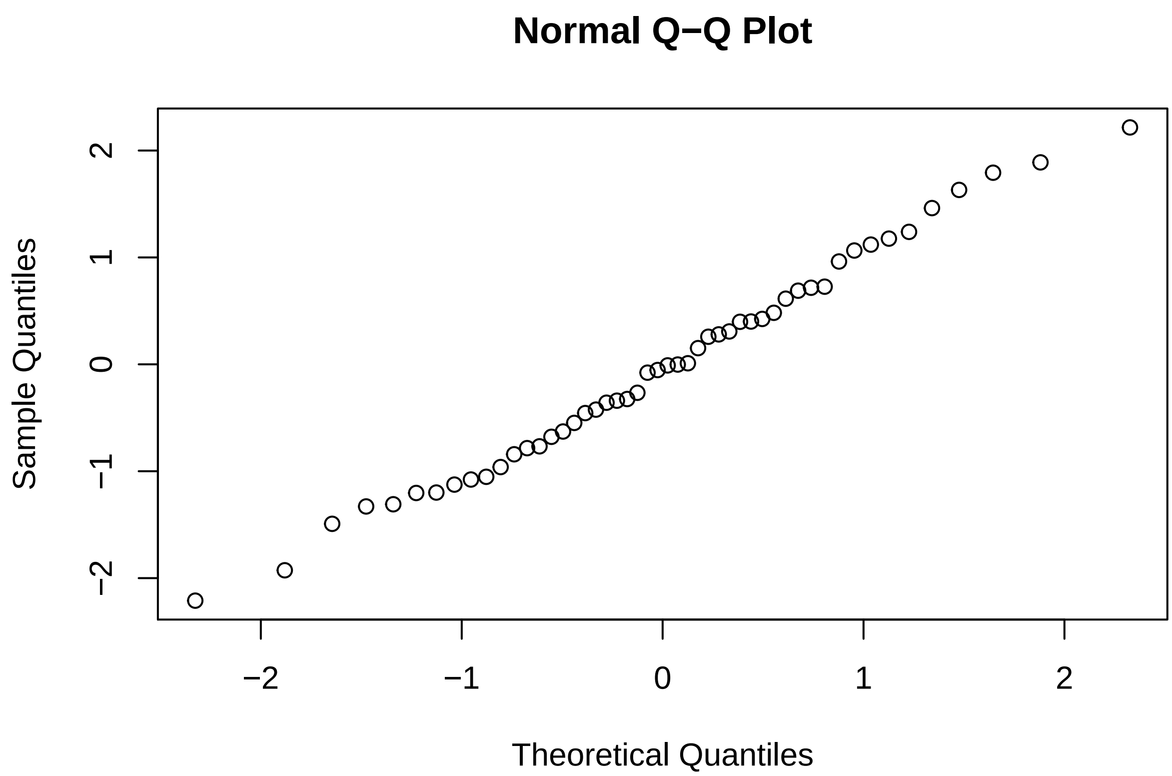
```
##
##  One-sample Kolmogorov-Smirnov test
##
```

```
## data:  m4
## D = 0.06722, p-value = 0.9663
## alternative hypothesis: two-sided
```

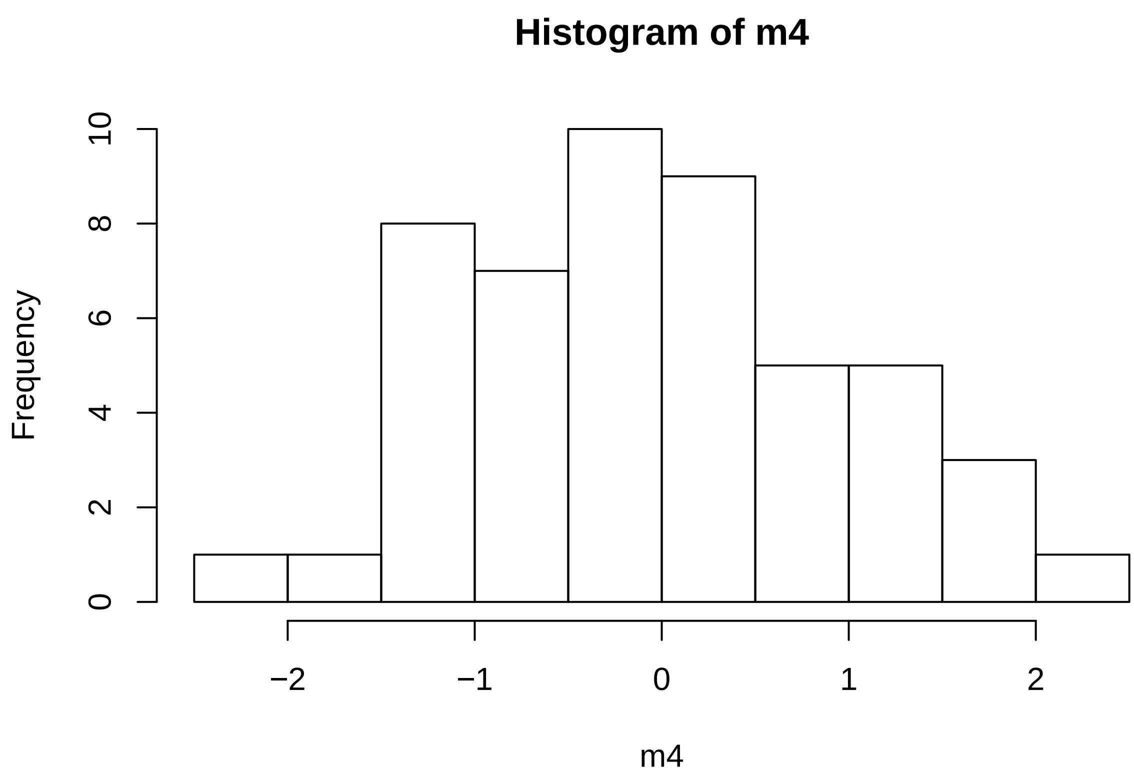
```r
plot2 <- ecdf(m4)
plot(plot2)
```




```r
qqnorm(m4)
```




```r
hist(m4)
```



From the ecdf plot as well as the ks-test, we may conclude that it is almost a normal distribution.

Question 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_same_1 <- read.table("maybe_same_1.txt")
```

```
## Warning in read.table("maybe_same_1.txt"): incomplete final line found by
## readTableHeader on 'maybe_same_1.txt'
```

```
maybe_same_2 <- read.table("maybe_same_2.txt")
```

```
## Warning in read.table("maybe_same_2.txt"): incomplete final line found by
## readTableHeader on 'maybe_same_2.txt'
```

```
maybe_same_1 <- c(maybe_same_1$V1, maybe_same_1$V2, maybe_same_1$V3, maybe_same_1$V4,maybe_same_1$V5)
maybe_same_2 <- c(maybe_same_2$V1, maybe_same_2$V2, maybe_same_2$V3, maybe_same_2$V4,maybe_same_2$V5)
maybe_same <- c(maybe_same_1,maybe_same_2)
ks.test(maybe_same_1, maybe_same_2)
```

```
## Warning in ks.test(maybe_same_1, maybe_same_2): cannot compute exact p-
## value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  maybe_same_1 and maybe_same_2
## D = 0.25, p-value = 0.491
## alternative hypothesis: two-sided
```

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

```
## Warning in ks.test(maybe_same_1, maybe_same): cannot compute exact p-value
## with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  maybe_same_1 and maybe_same
## D = 0.11111, p-value = 0.9888
## alternative hypothesis: two-sided
```

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

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

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  maybe_same_2 and maybe_same
## D = 0.13889, p-value = 0.9522
## alternative hypothesis: two-sided
```

```
maybe_same_1_add <- maybe_same_1 + 2
ks.test(maybe_same_1_add, maybe_same_2)
```

```
## Warning in ks.test(maybe_same_1_add, maybe_same_2): cannot compute exact p-
## value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  maybe_same_1_add and maybe_same_2
## D = 0.65, p-value = 0.0001673
## alternative hypothesis: two-sided
```

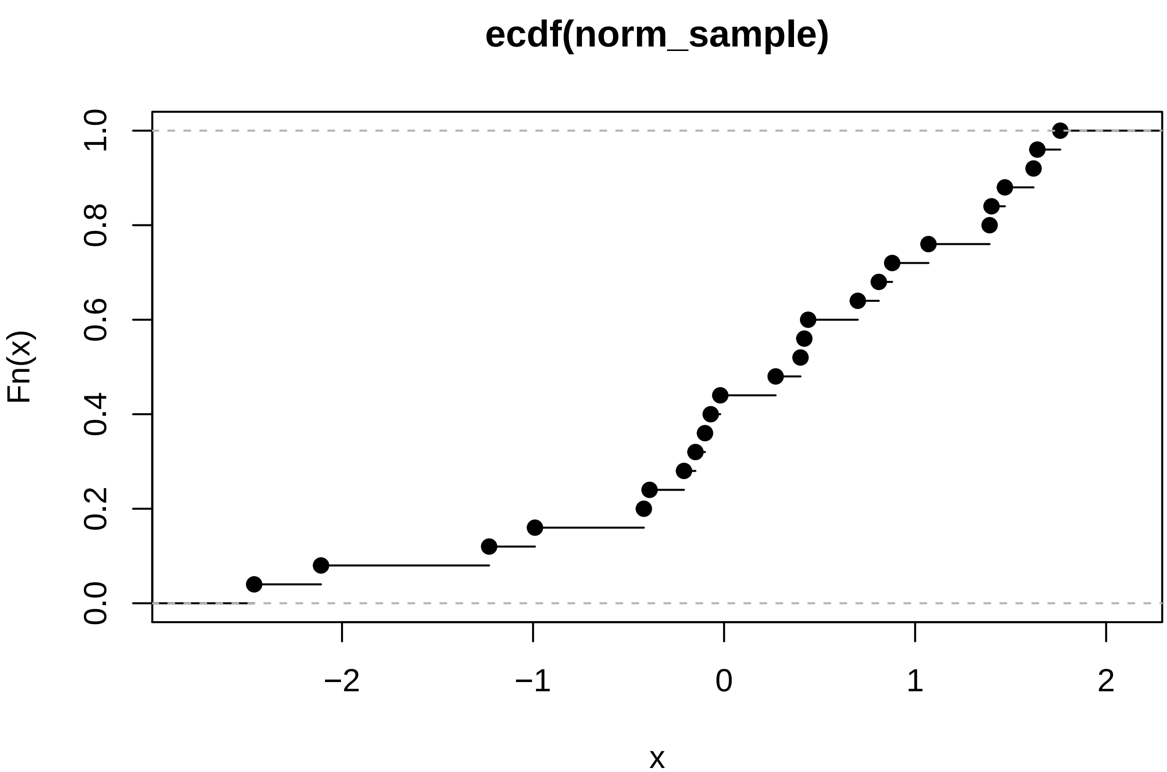
We fount that the p value of ks test between x and y is 0.49, which is much larger than 0.05. Also, the p value of ks test between x and y+x is 0.98, p value of ks test between x+y and y is 0.95. So we have the evidence to say x and y are from the sme distribution.

On the other hand, p value of ks test between x+2 and y is 0.00017, which is much less than 0.05, so they are not from the same distribution.

Question4

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 Kolmogorov-Smirnov test. Calculate the D statistic. Run the ks.test() function and compare your results to the results reported by ks.test.

```r
norm_sample <- readRDS("norm_sample.Rdata")
q4 <- ecdf(norm_sample)
plot(q4)
```




```r
test <- rnorm(n = 25, 0, 1)

ks.test(test,norm_sample)
```

```
##
```

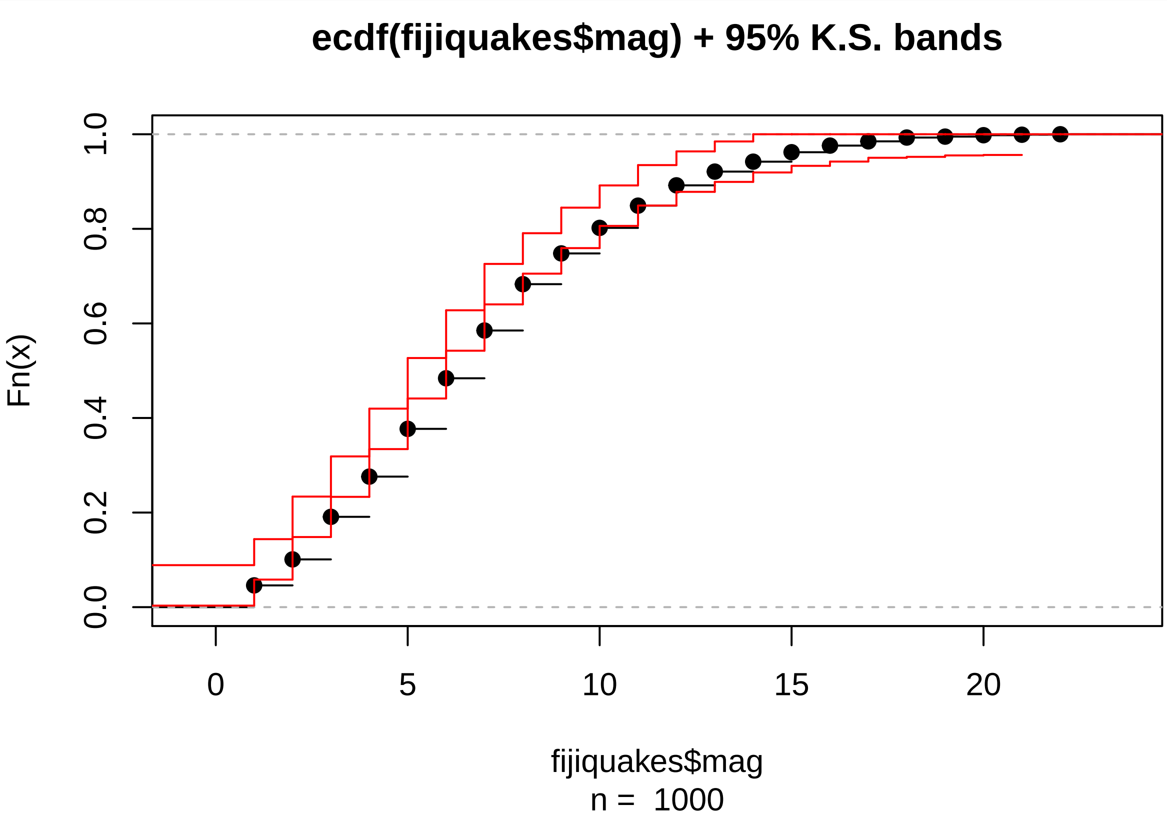
```
##  Two-sample Kolmogorov-Smirnov test
##
## data:  test and norm_sample
## D = 0.16, p-value = 0.915
## alternative hypothesis: two-sided
```

Here we find that D=0.2, and p-value is 0.71, which might help me conclude that they are the same distribution. But on the other hand, the edcf plot doesn't show a pretty strong evidence of same distribution. The plot is becoming more precipitous from -2 to 2.

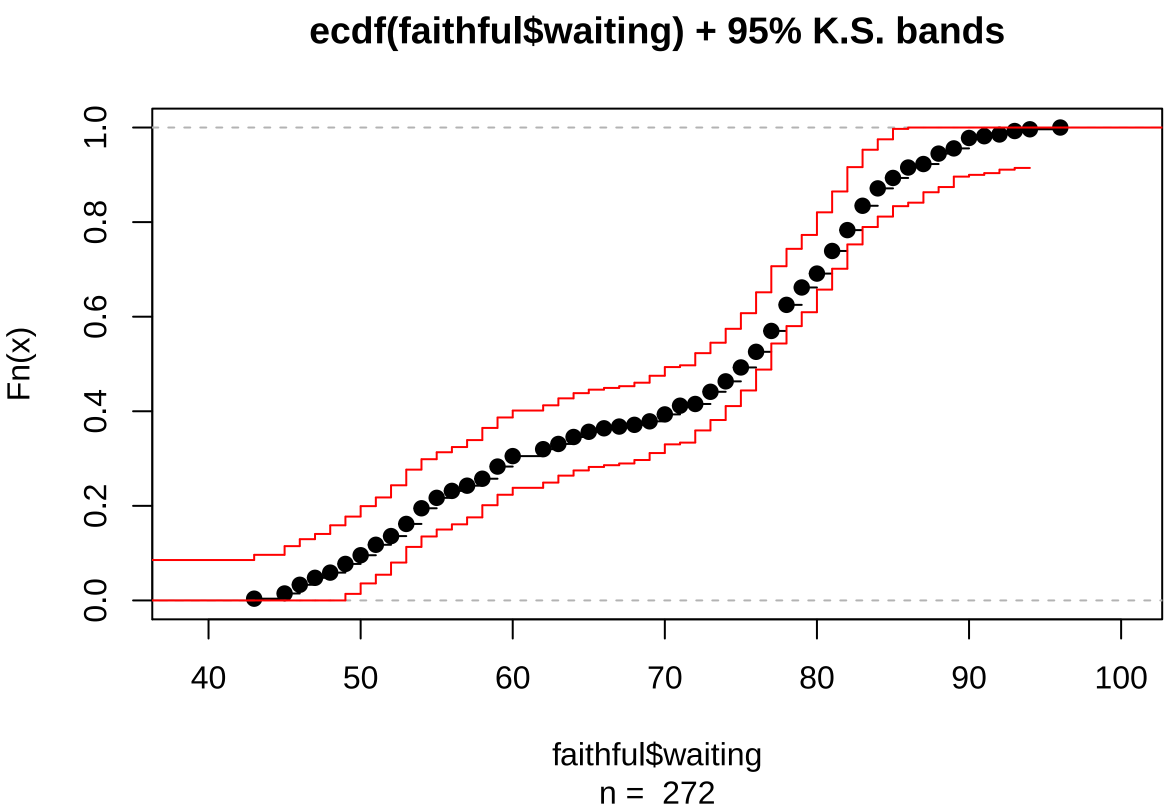
Question 5

Produce empirical distributions with confidence bands for the fuji-quakes.dat and faithful.dat. For the fujiquakes data, Find a 95 for 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.

```
fijiquakes <- read.table("fijiquakes.dat" )
index <- fijiquakes[1,]
fijiquakes <- fijiquakes[-1,]
colnames(fijiquakes) <- c("Obs","lat","long","depth","mag","stations")

faithful <- read.table("faithful.dat", skip = 25)

q5_1 <- ecdf(fijiquakes$mag)
q5_2 <- ecdf(faithful$waiting)


ecdf.ksCI(fijiquakes$mag)
```



```r
ecdf.ksCI(faithful$waiting)
```




```r
#mean
faith_mean = mean(faithful$waiting)
faith_se = sd(faithful$waiting)/sqrt(length(faithful$waiting))

t1 <- faith_mean - 0.6*faith_se
t2 <- faith_mean + 0.6*faith_se
t1
```

```
## [1] 70.40247
```

```r
t2
```

```
## [1] 71.39165
```

```r
#median:
summary(faithful)
```

```
##    eruptions        waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

Thus, we can know that the interval for mean waiting time is [70.40247,71.39165]. the median waitint time is 76.