Lab 01

Amanda Mentesana

2025-01-24

Reading the data set and filtering out NA values.

```
EPI_data <- read.csv("C:/Users/amanda/Downloads/epi2024results06022024.csv")

attach(EPI_data) #sets as default object, can call variables directly such as EPI_data$EPI.new is EPI.n

#head(EPI_data) #to show first part of data frame

# records True values if the value is NA
NAs <- is.na(EPI.new)
# filters out NA values, new array
EPI.new.noNAs <- EPI.new[!NAs]
```

Excercise 1: Exploring the Distribution

```
summary(EPI_data$EPI.new) #Output shows Median, Mean, Quartiles and range
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.50   38.25   45.50   46.84   53.10   75.30
```

```
fivenum(EPI.new,na.rm=TRUE) #Output shows the same values as summary() without heading
```

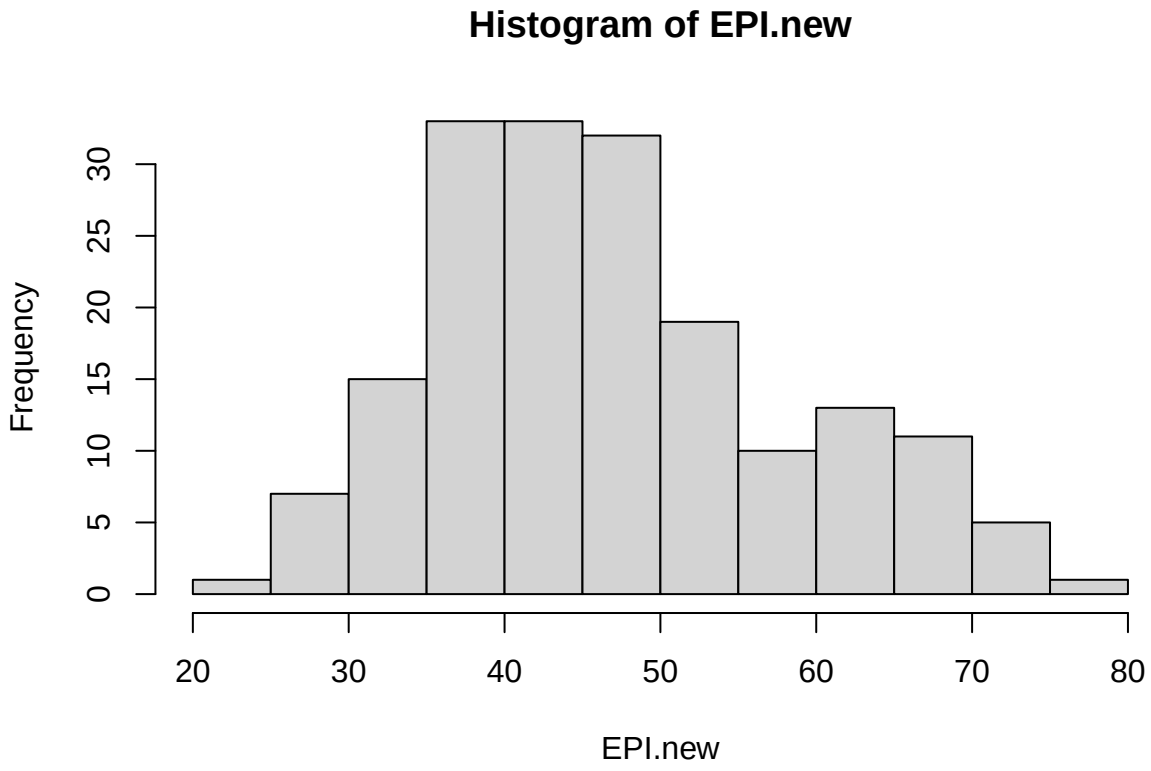
```
## [1] 24.5 38.2 45.5 53.1 75.3
```

```
stem(EPI.new) # stem and leaf plot
```

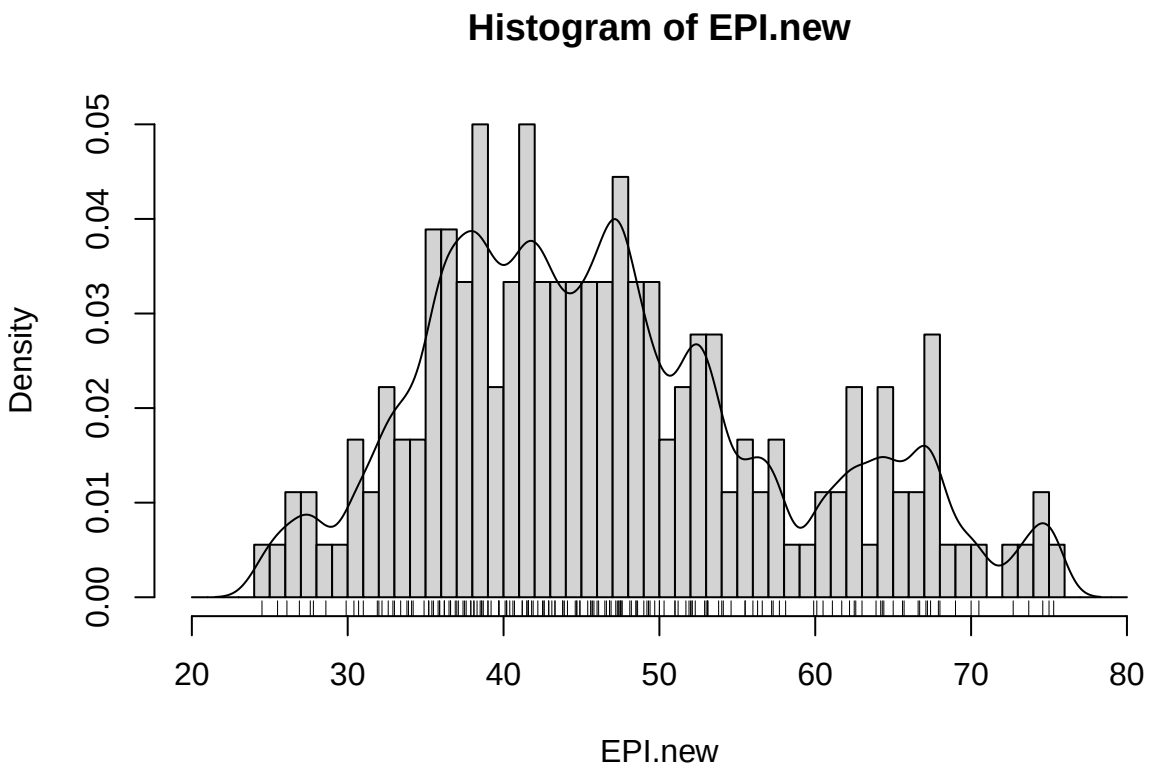
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   2 |
##   2 | 5667889
##   3 | 001122233334444
##   3 | 55556666667777777888888899999999
##   4 | 00000001111222222233333334444
##   4 | 555555666666777777777888888999999
##   5 | 000011122222233333444
##   5 | 5666677788
```

```
##   6 | 0011223334444
##   6 | 56677777889
##   7 | 0134
##   7 | 555
```

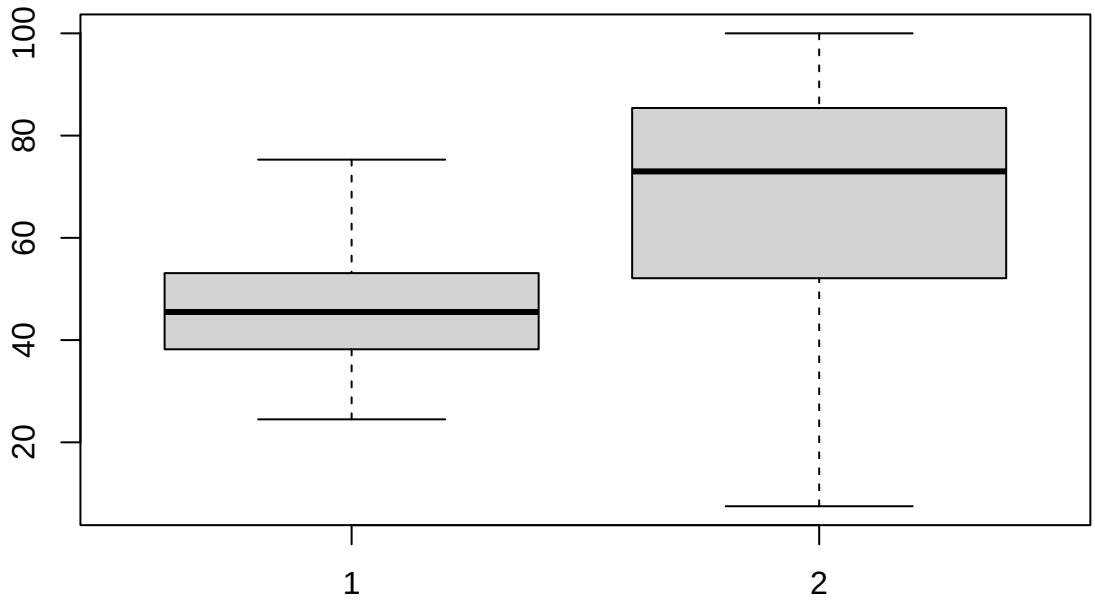
```
hist(EPI.new) #basic histogram with default bin sizes
```




```
hist(EPI.new, seq(20., 80., 1.0), prob=TRUE) #histogram with a set range (20 to 80) and bin size of 1

lines(density(EPI.new,na.rm=TRUE,bw=1.)) # adds density plot over histogram, or try bw="SJ" for a smoot

rug(EPI.new) #rug plot below graph
```

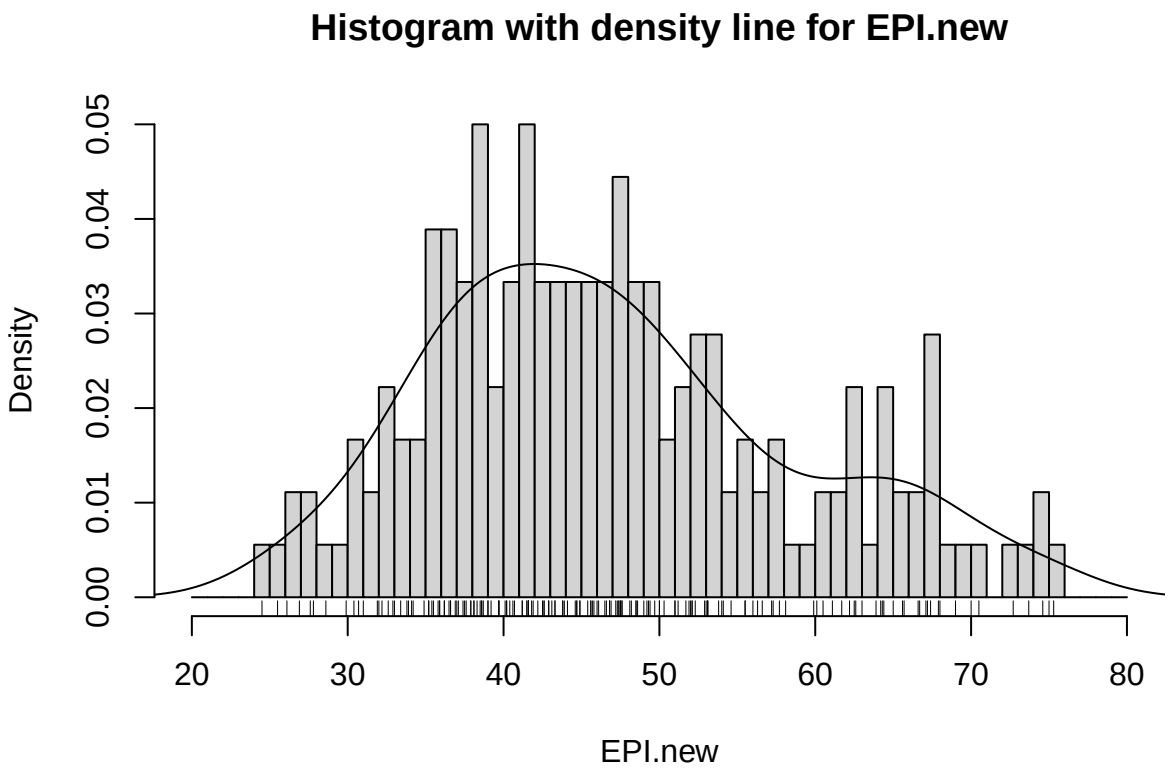



```
boxplot(EPI.new, APO.new) #using a boxplot to visually compare variables for EPI.new and APO.new
```



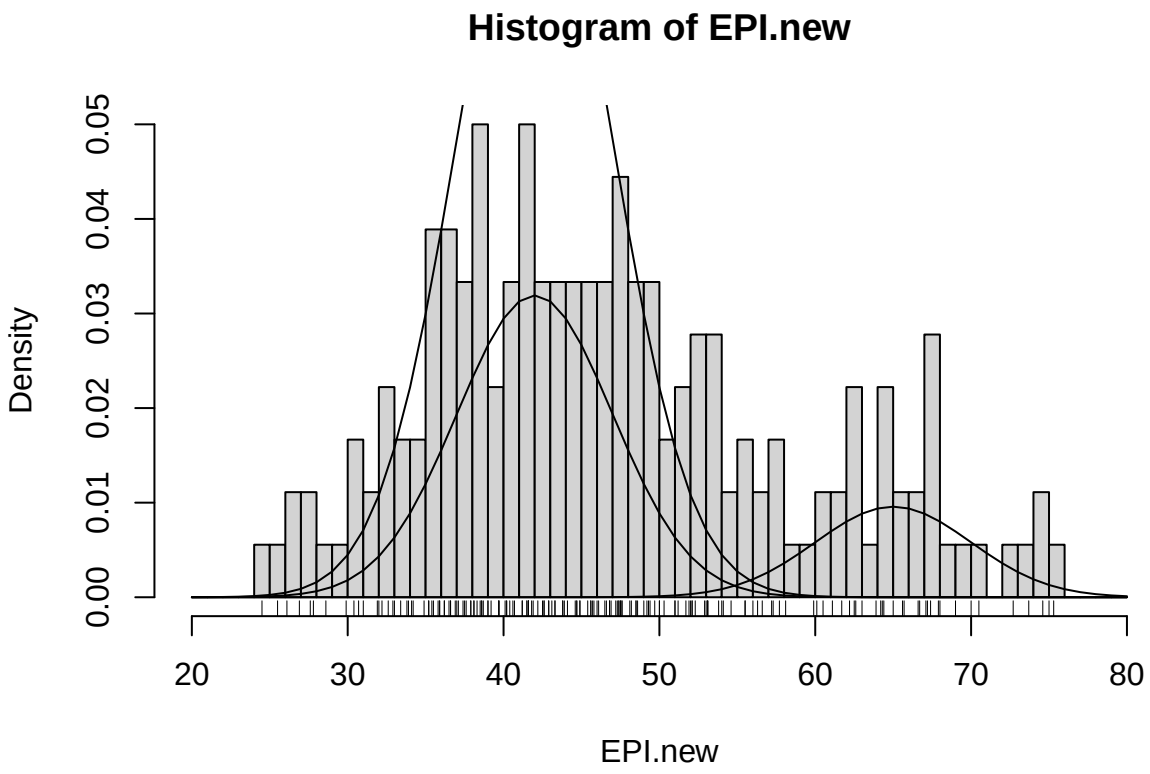
The first histogram using the kernel density estimation shows there are two areas where the data may be centered, at about 42 and 65.

```
#Histogram with kernal density estimation

hist(EPI.new, seq(20., 80., 1.0), main=paste("Histogram with density line for EPI.new"), prob=TRUE)

lines(density(EPI.new,na.rm=TRUE,bw="SJ"))

rug(EPI.new)
```



To further look at these trends, using the normal distribution, we can see there are trends in the data which show two populations at 42 and 65.

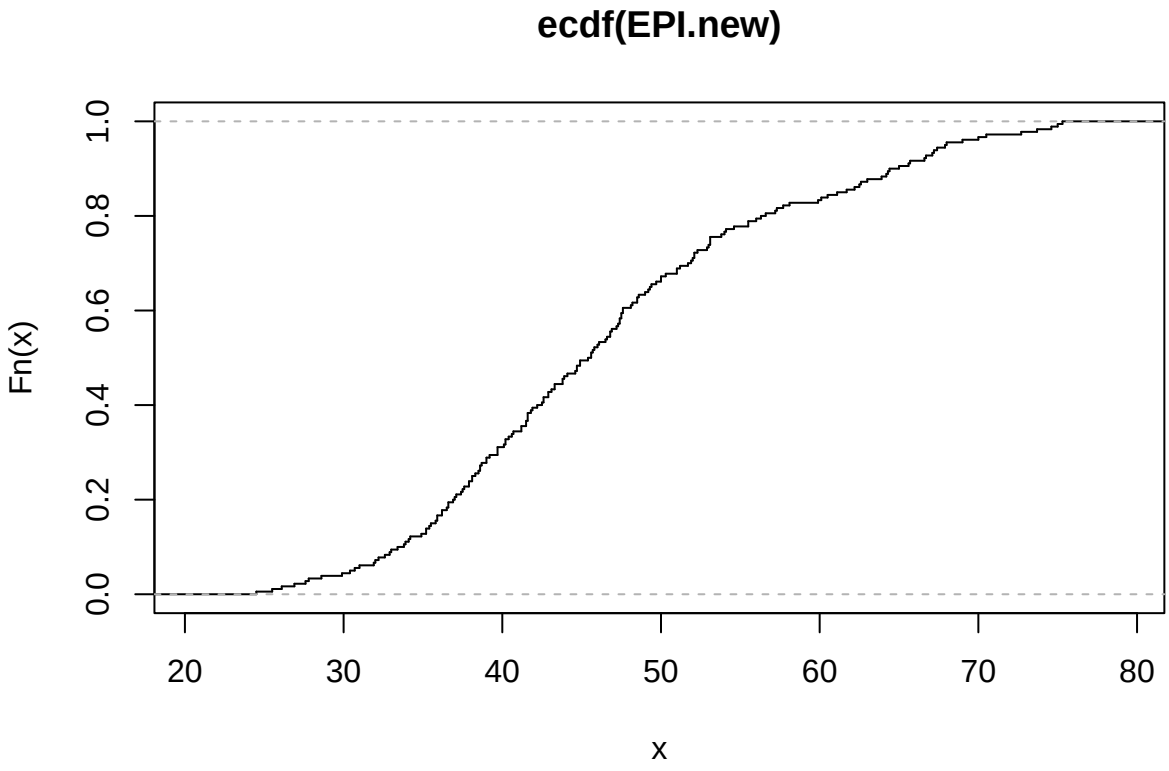
```
# histogram with normal distribution estimations
hist(EPI.new, seq(20., 80., 1.0, ),prob=TRUE)

#lines(density(EPI.new,na.rm=TRUE,bw="SJ"))

rug(EPI.new)

x<-seq(20,80,1)

q<- dnorm(x,mean=42, sd=5,log=FALSE) #fitting to a normal curve distribution with mean centered at 42 a

lines(x,q) #plots line q over this range

lines(x,.4*q) # plots 40% of q over the range

p<-dnorm(x,mean=65, sd=5,log=FALSE) #fitting to a normal curve distribution with mean centered at 65 an

lines(x,.12*p) # plots 12% of p over the given range
```



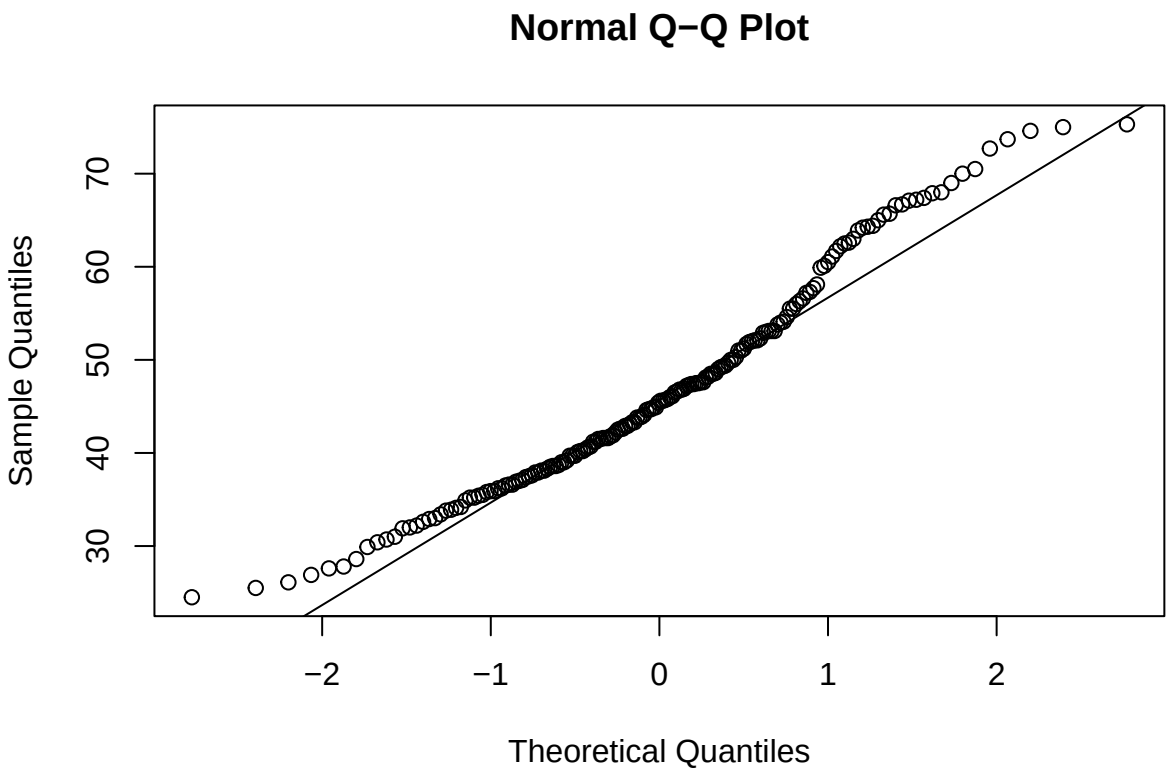
Excercise 2: Fitting a Distribution

Based on the distributions, the EPI.new variable reasonably fits the normal distribution fit, and shows possible bimodality.

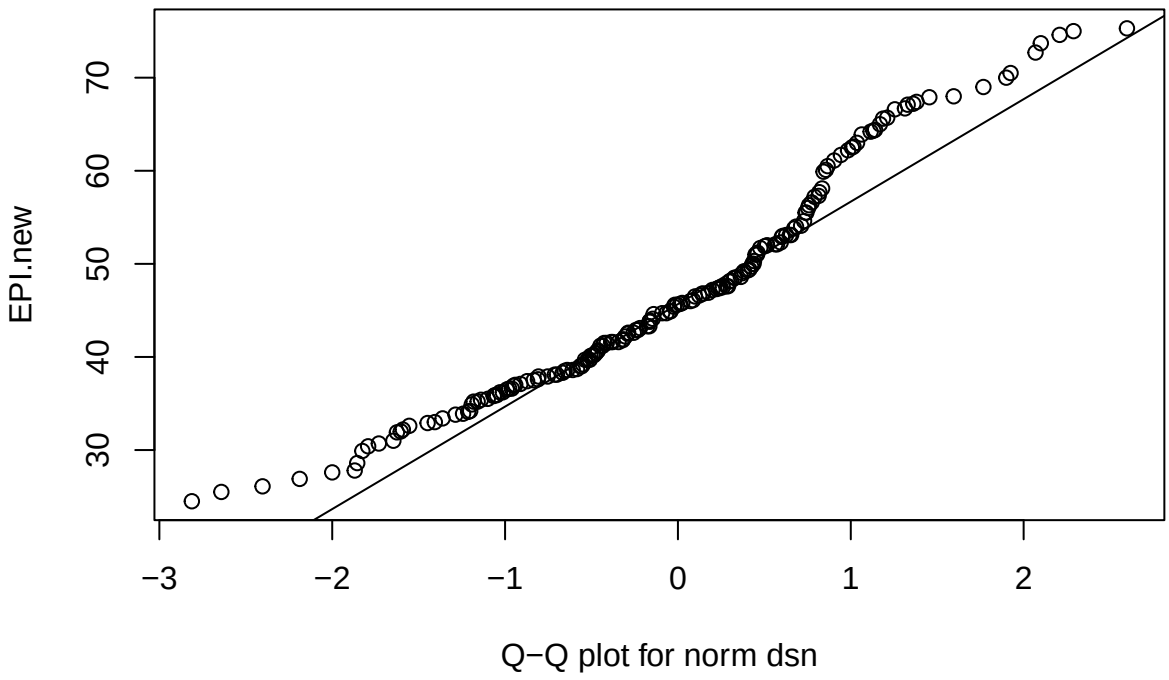
```
plot(ecdf(EPI.new), do.points=FALSE, verticals=TRUE)
```



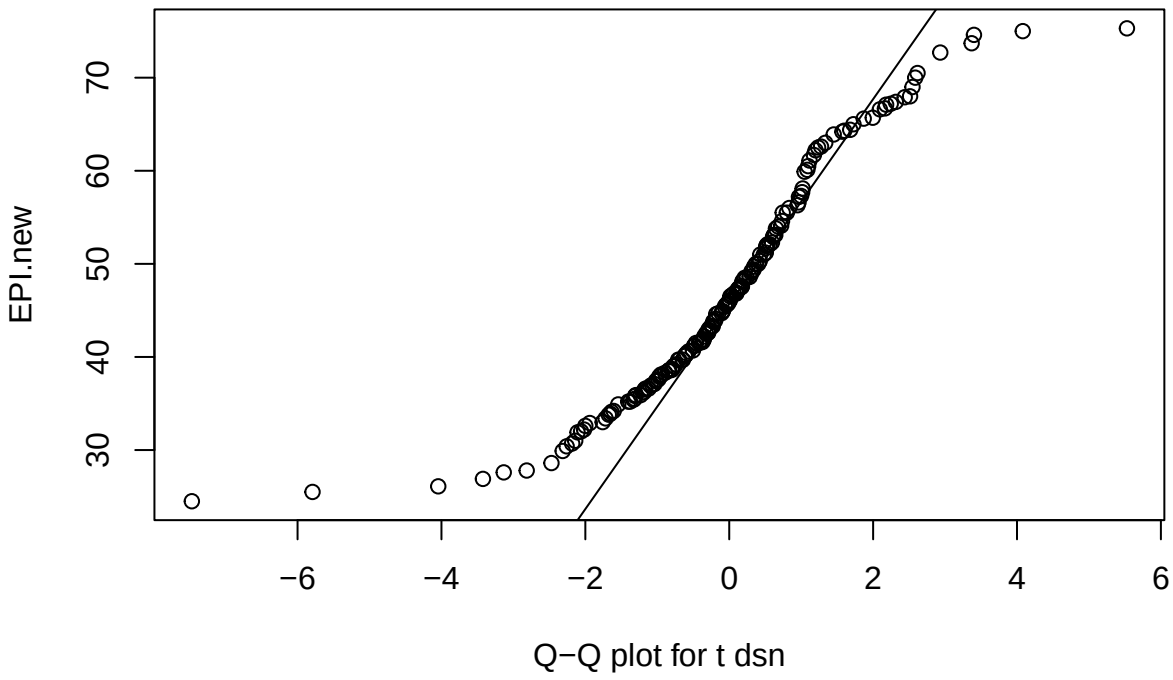

```
qqnorm(EPI.new) #a standard quantile-quantile plot of EPI dataset
qqline(EPI.new) #baseline
```




```
qqplot(rnorm(250), EPI.new, xlab = "Q-Q plot for norm dsn") #qq plot fit for the normal distribution
qqline(EPI.new)
```




```
qqplot(rt(250, df = 5), EPI.new, xlab = "Q-Q plot for t dsn") #qq plot plot fit to the t-dstribution wi
qqline(EPI.new)
```



Excercise 2a: Fitting the MPE Variable

We can see that the MPE variable has a different distribution and range than EPI. The mean is 69.85, yet both the 3rd quartile and maximum value is 100. When looking at the stem-and-leaf plot we can see that the data is skewed towards the value 100.

```
summary(MPE.new) #Output shows Median, Mean, Quartiles and range
```

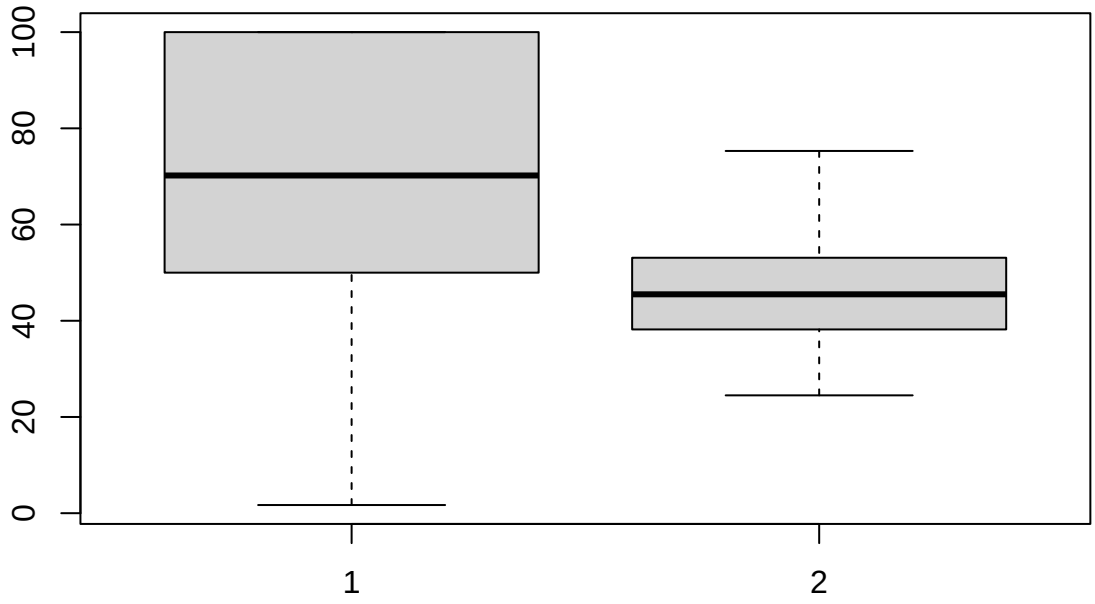
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.70   50.00   70.20   69.85  100.00  100.00      49
```

```
stem(MPE.new) # stem and leaf plot of SPI
```

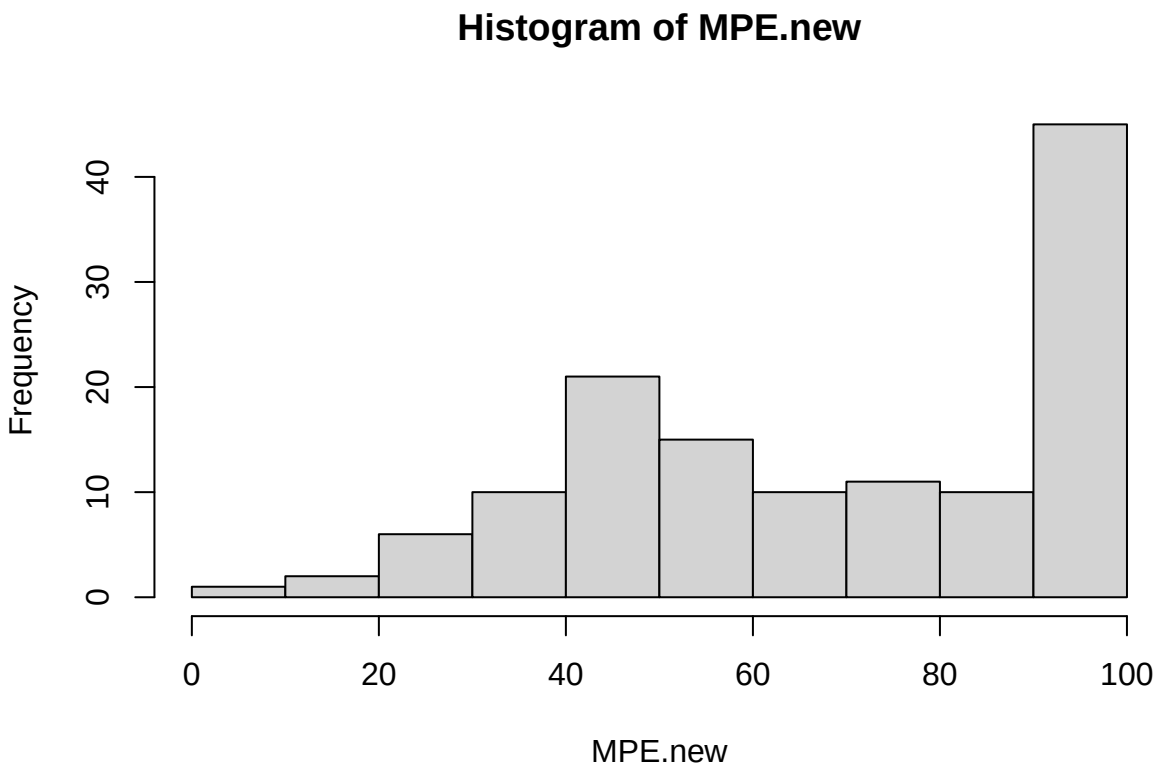
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 2
##    0 |
##    1 | 11
##    1 |
##    2 | 11
##    2 | 6677
##    3 | 13333
##    3 | 6777
```

```
##     4 | 01344
##     4 | 55557789
##     5 | 000000000133444
##     5 | 55557789
##     6 | 0111224
##     6 | 6799
##     7 | 00033
##     7 | 566788
##     8 | 4
##     8 | 566888999
##     9 | 1
##     9 | 67788
##    10 | 0000000000000000000000000000000000000000
```

```
boxplot(MPE.new, EPI.new) #using a box plot to visually compare variables for MPE.new and EPI.new
```

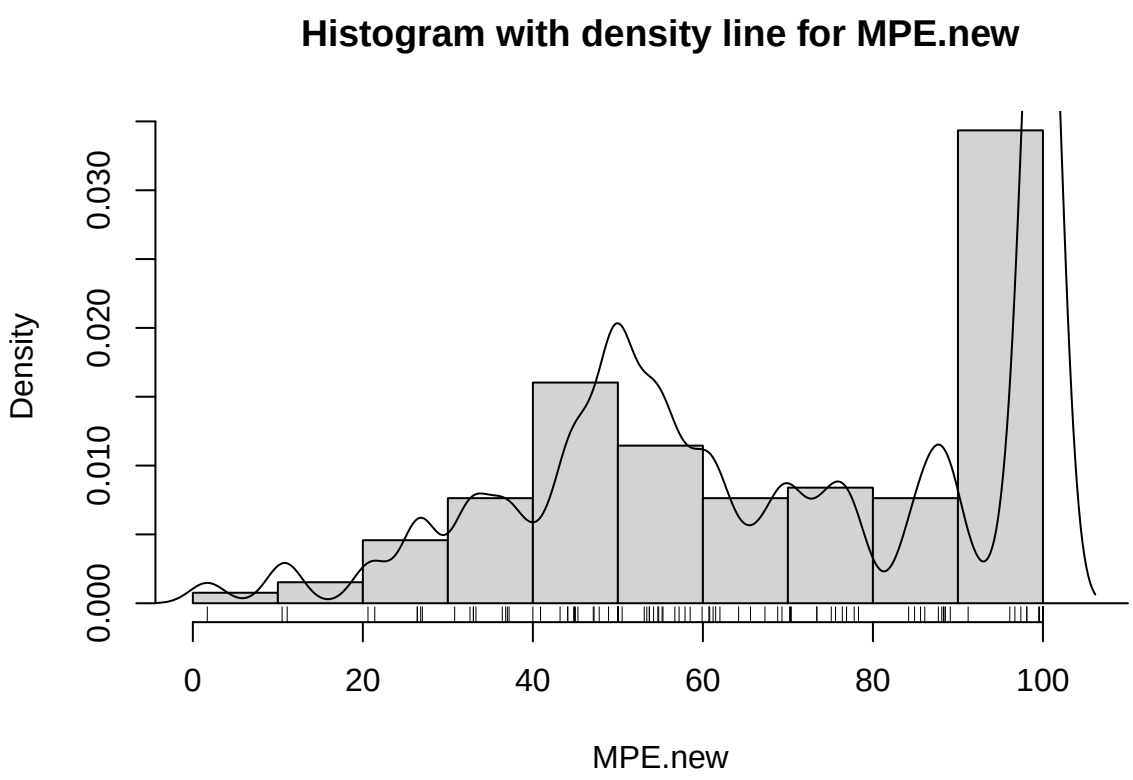



```
hist(MPE.new)
```



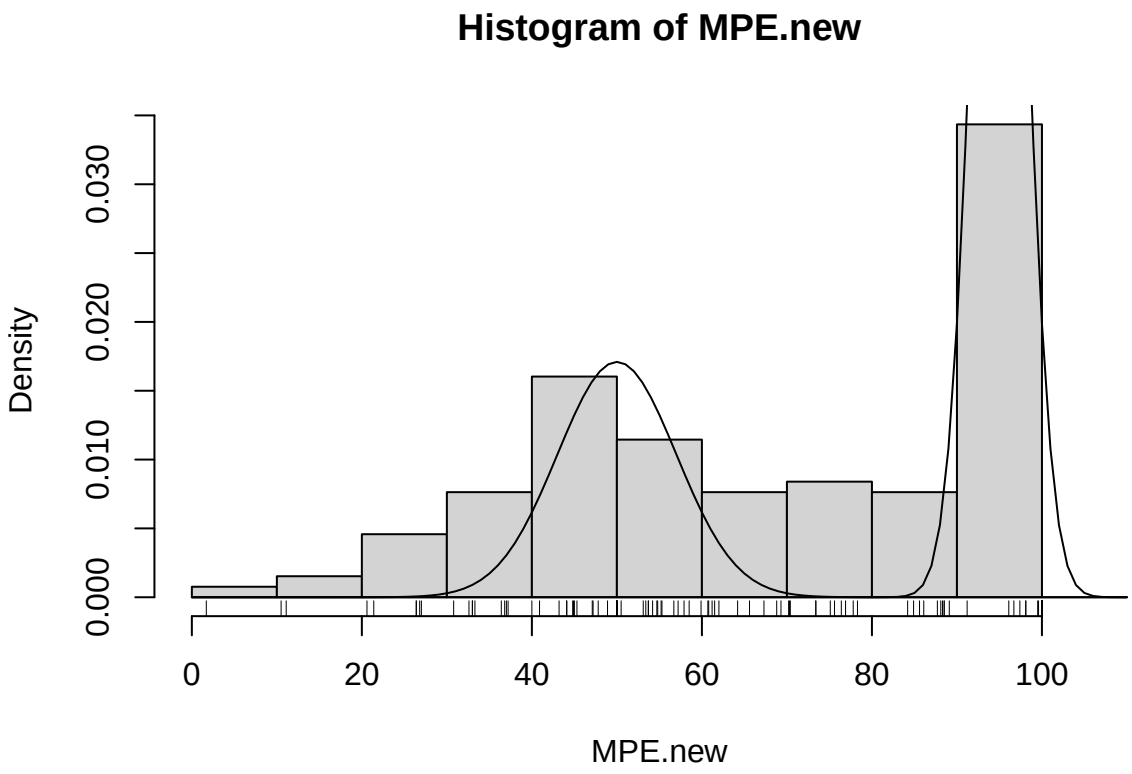
The first histogram using the kernal density estimation shows two peaks, with skewedness to 100.

```
#Histogram with kernal density estimation

hist(MPE.new, seq(0., 110., 10.), main=paste("Histogram with density line for MPE.new"), prob=TRUE)

lines(density(MPE.new,na.rm=TRUE,bw="SJ"))

rug(MPE.new)
```



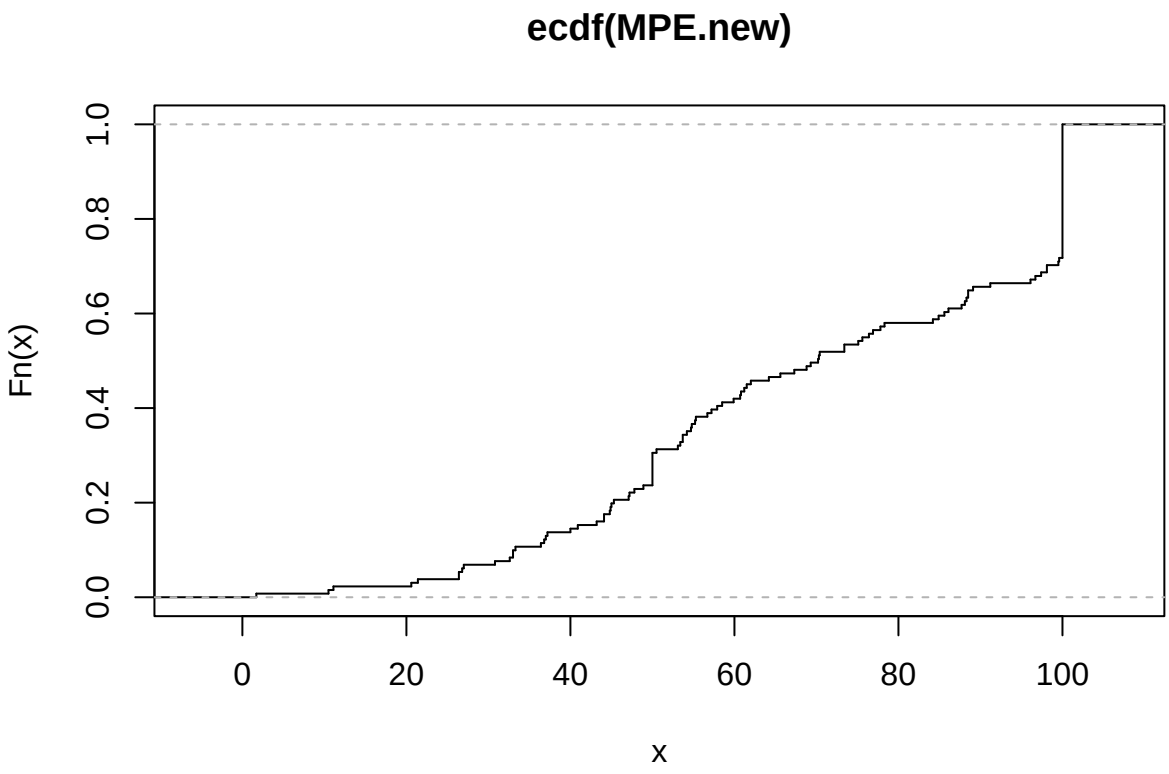
To further look at these trends, using the normal distribution, we can see there are trends in the data which show two populations of data, one centered around 50 with a large spread of data and the other at 100 with a very narrow range of data.

```
# histogram with normal distribution estimations
hist(MPE.new, seq(0., 110., 10., ),prob=TRUE)

#lines(density(EPI.new,na.rm=TRUE,bw="SJ"))

rug(MPE.new)

x<-seq(0,110,1)

q<- dnorm(x,mean=50, sd=7,log=FALSE) #fitting to a normal curve distribution with mean centered at 50 a

lines(x,.3*q) # plots 0.3*q over the range

p<-dnorm(x,mean=95, sd=3,log=FALSE) #fitting to a normal curve distribution with mean centered at 95 an

lines(x,.6*p) # plots 0.6*p over the given range
```

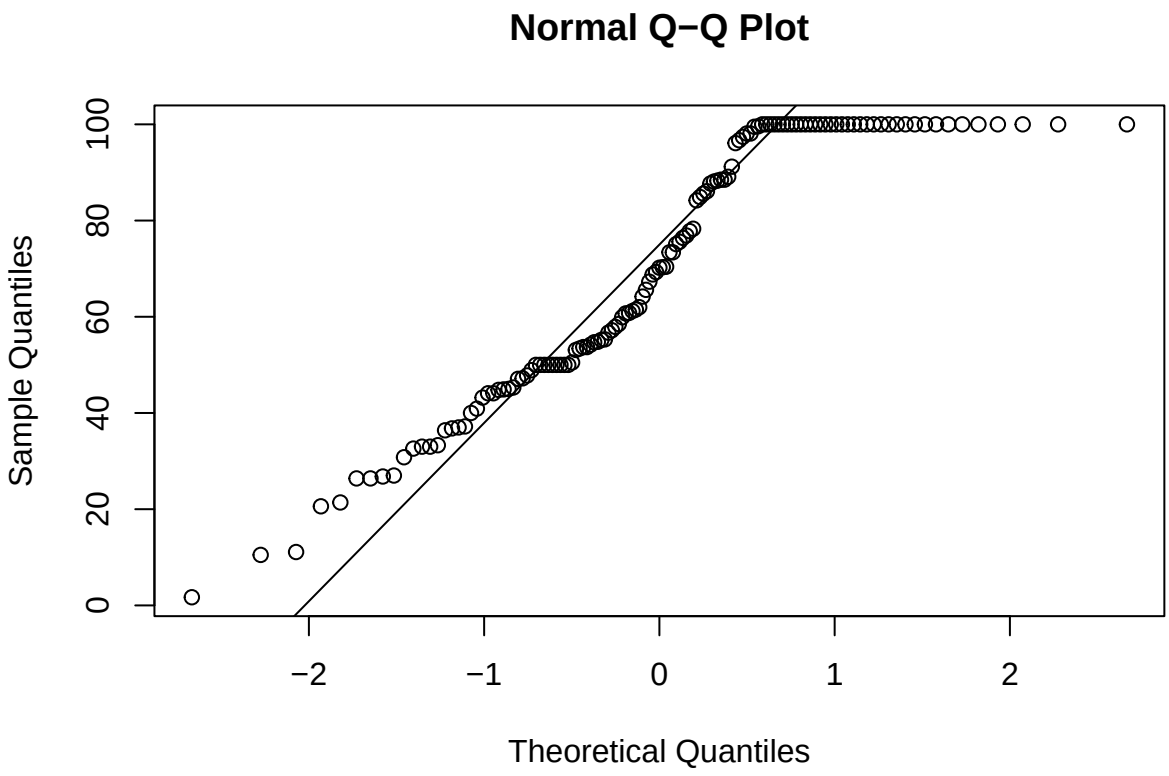


The normal Q-Q plot shows that data is too peaked at the end of the distribution to fit it well.

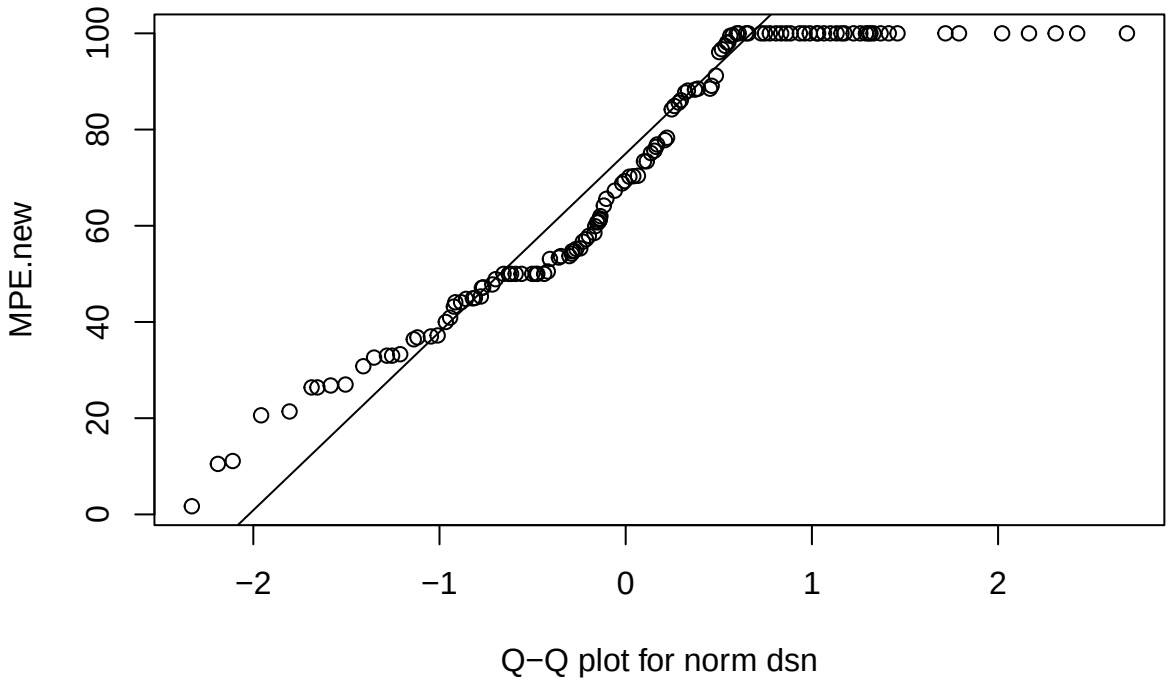
```
plot(ecdf(MPE.new), do.points=FALSE, verticals=TRUE)
```



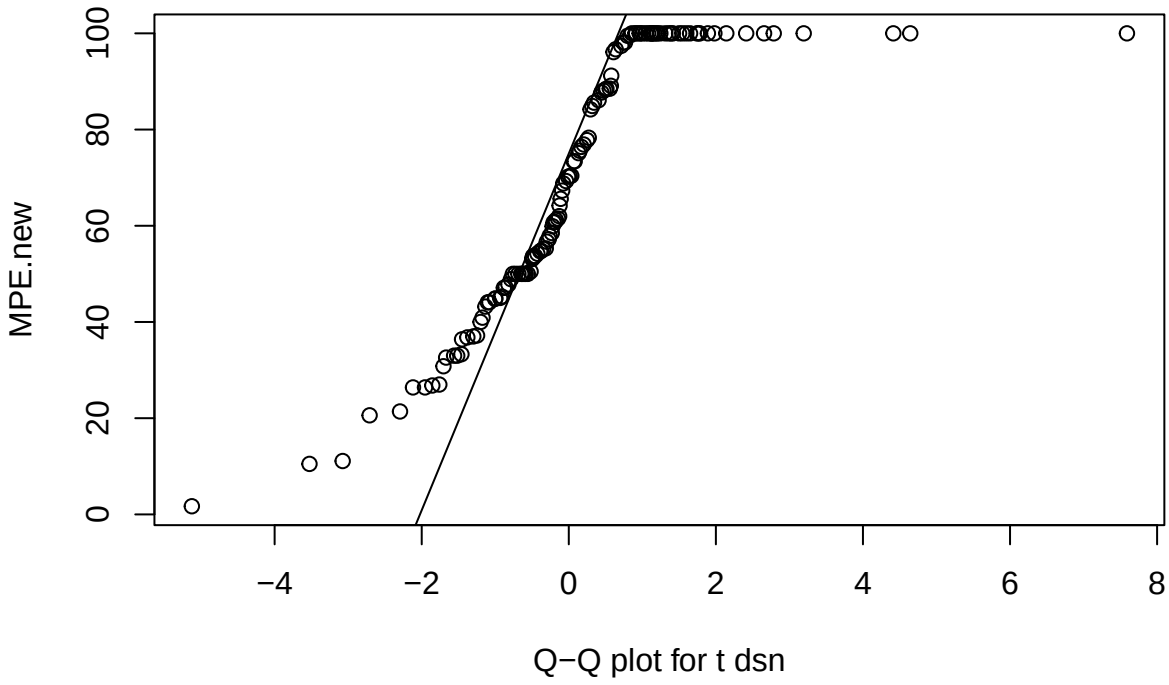

```
qqnorm(MPE.new); qqline(MPE.new)
```




```
qqplot(rnorm(250), MPE.new, xlab = "Q-Q plot for norm dsn")
qqline(MPE.new)
```




```
qqplot(rt(250, df = 5), MPE.new, xlab = "Q-Q plot for t dsn")
qqline(MPE.new)
```



Excercise 2b: Fitting the SPI Variable

We can see that the SPI variable has a larger spread than the EPI variable. The data also seems equally distributed using the stem and leaf plot.

```
summary(SPI.new) #Output shows Median, Mean, Quartiles and range
```

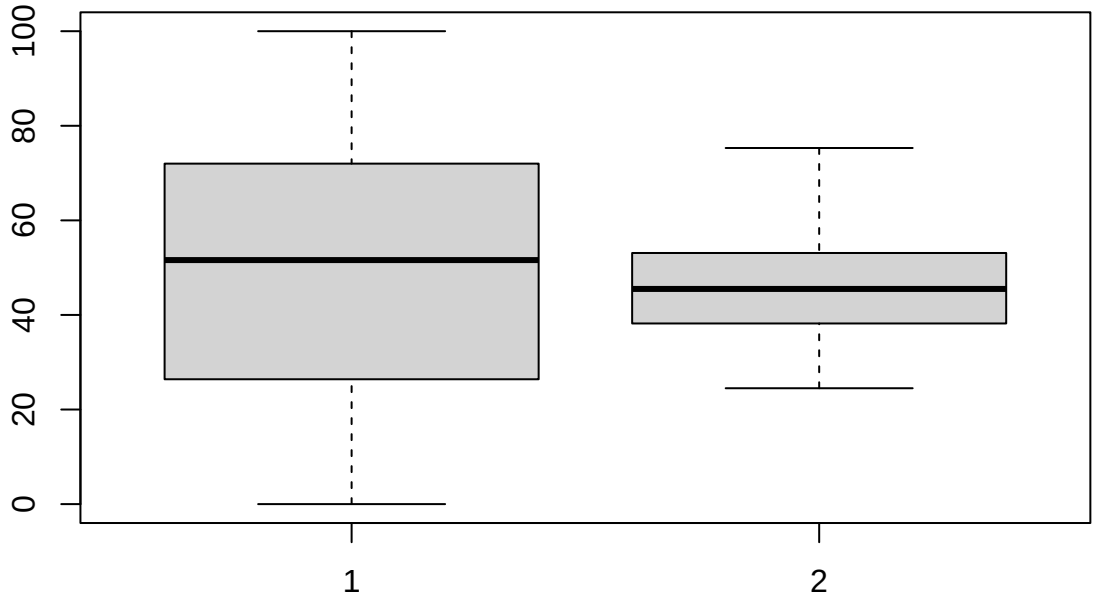
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   26.88   51.60   49.84   71.95  100.00       2
```

```
#fivenum(SPI.new,na.rm=TRUE) #Output shows the same values as summary() without heading

stem(SPI.new) # stem and leaf plot of SPI
```

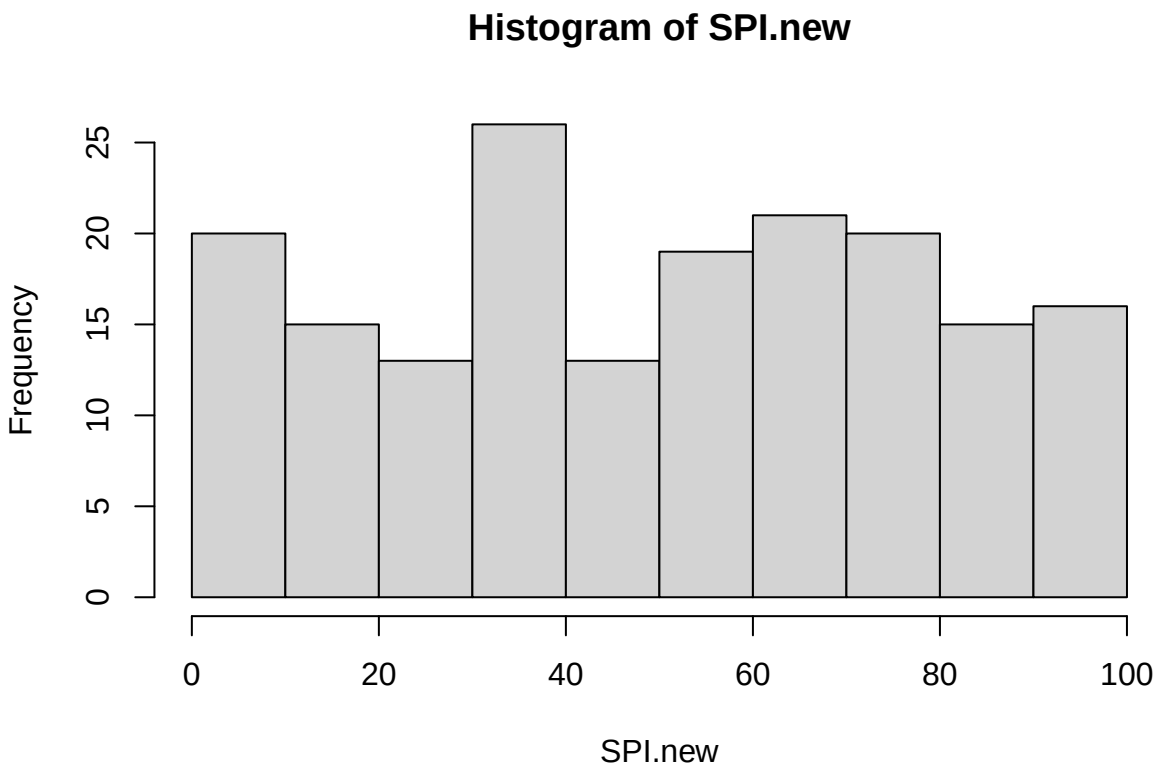
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 00011113444445788889
##    1 | 1344455788999
##    2 | 001122355566889
##    3 | 11223334445556777777789999
##    4 | 0122333356779
##    5 | 00123444567778888899
##    6 | 0233355566788889999
```

```
##    7 | 0000122222334556788999
##    8 | 1123344455579
##    9 | 00011124445666789
##   10 | 0
```

```r
boxplot(SPI.new, EPI.new) #using a boxplot to visually compare variables for SPI.new and EPI.new
```

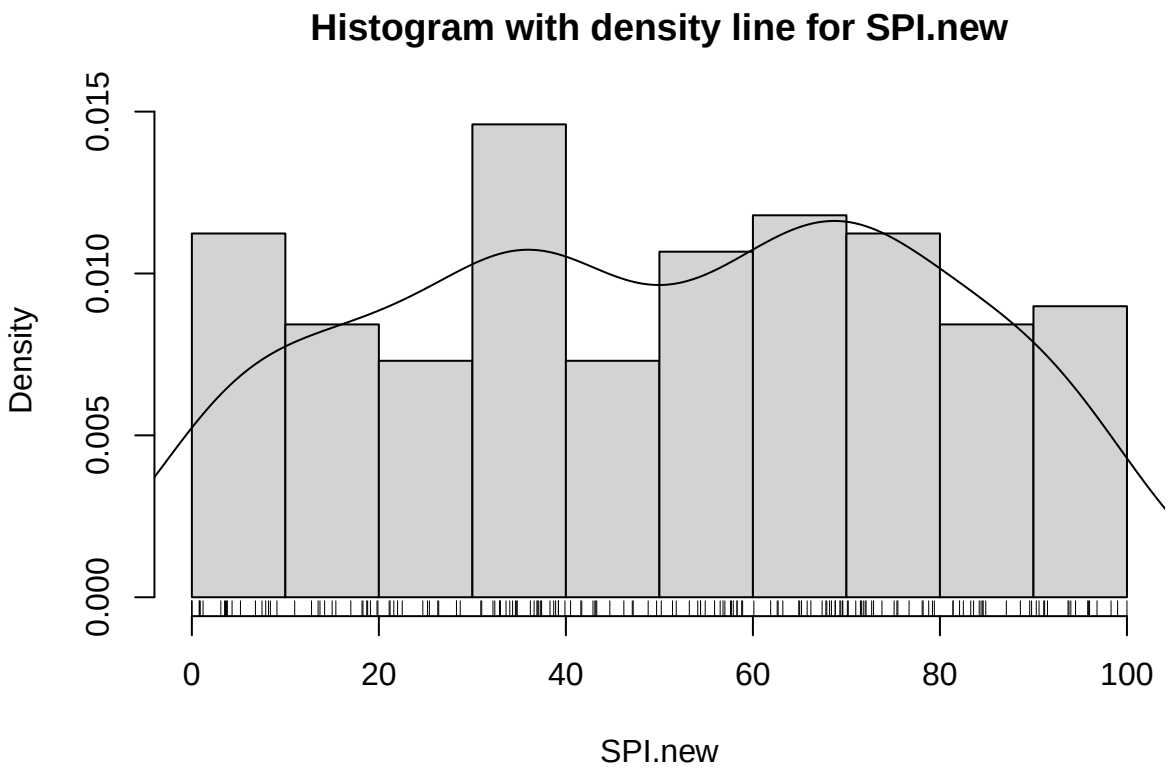



```r
hist(SPI.new)
```



The first histogram using the kernal density estimation shows very low variation in the SPI.new data distribution. Fitting to a normal curve was not helpful in describing the data.

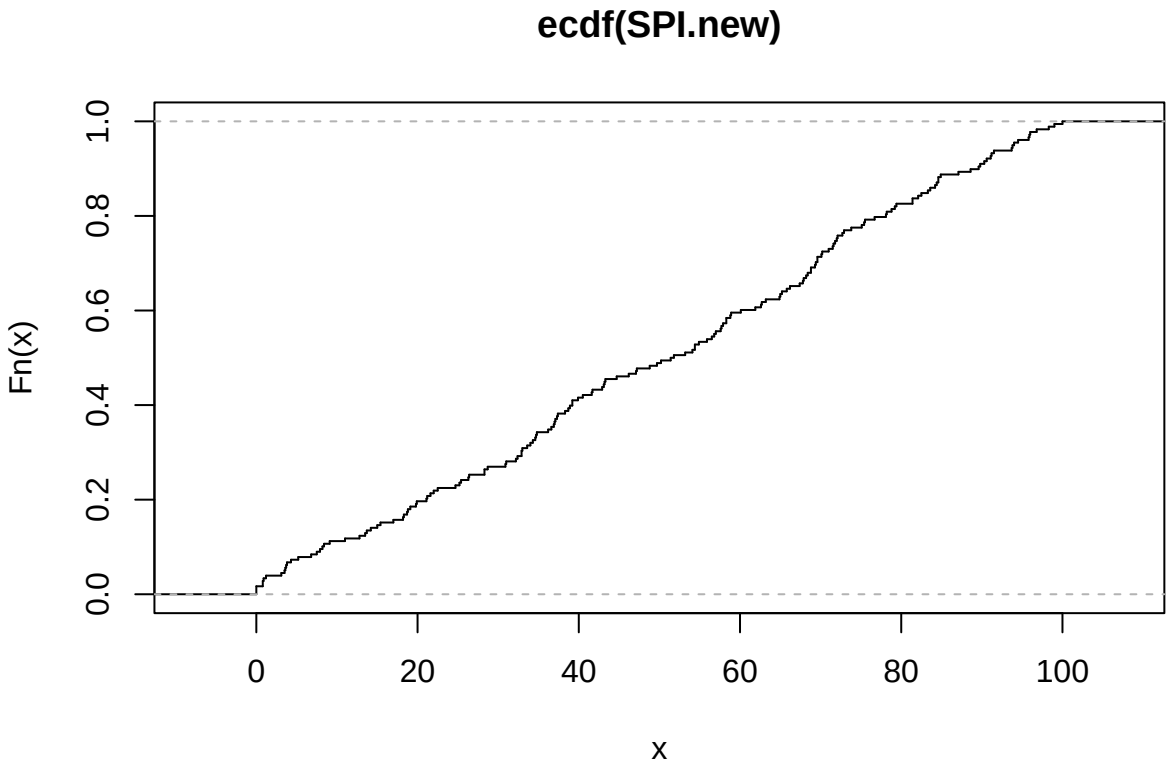
```
#Histogram with kernal density estimation

hist(SPI.new, seq(0., 100., 10), main=paste("Histogram with density line for SPI.new"), prob=TRUE)
lines(density(SPI.new,na.rm=TRUE,bw="SJ"))
rug(SPI.new)
```



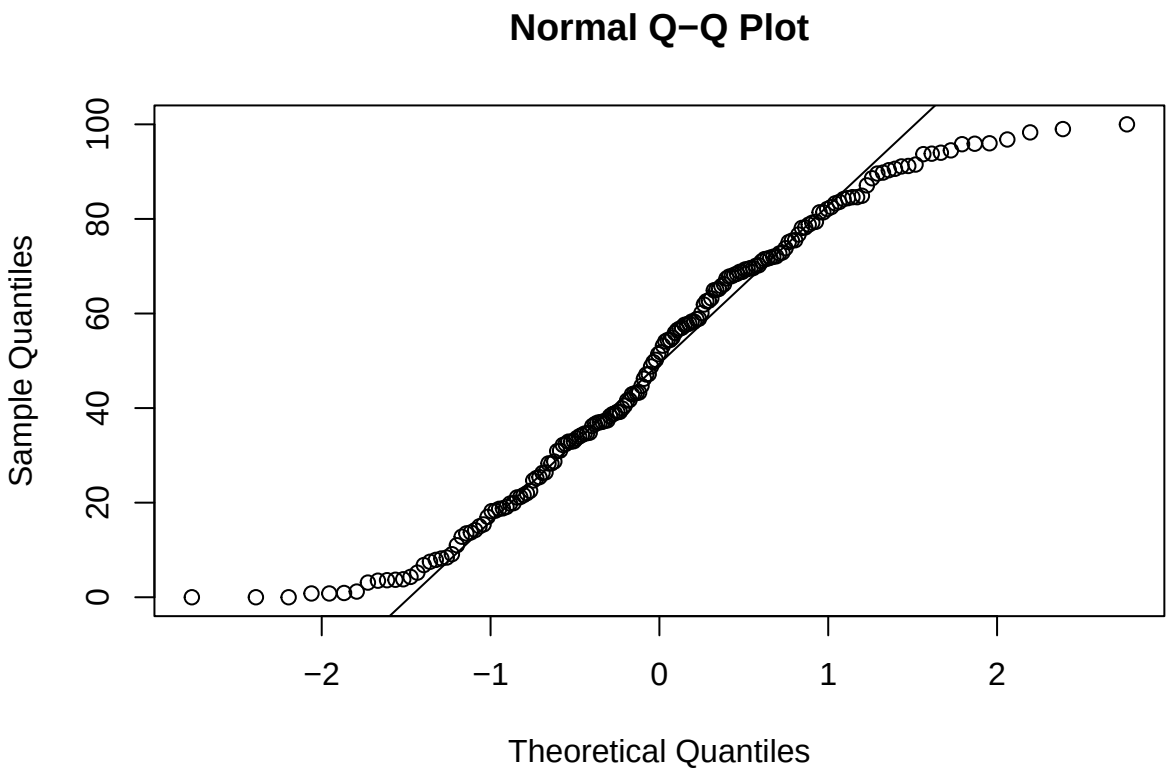

```
x<-seq(0,100,10)

q<- dnorm(x,mean=50, sd=3,log=FALSE) #fitting to a normal curve distribution with mean centered at 42 a

#lines(x,q) #plots line q over this range
```

The normal Q-Q plot shows the SPI.new data is heavily tailed however it is fit to the data well.

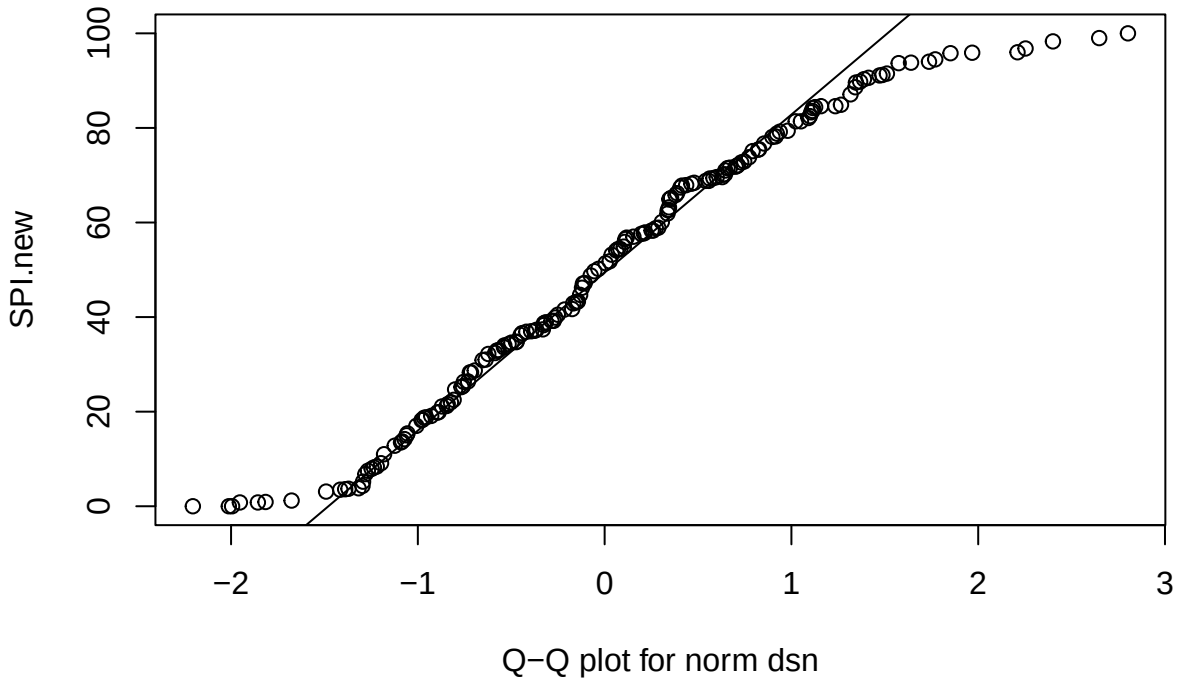
```
plot(ecdf(SPI.new), do.points=FALSE, verticals=TRUE)
```



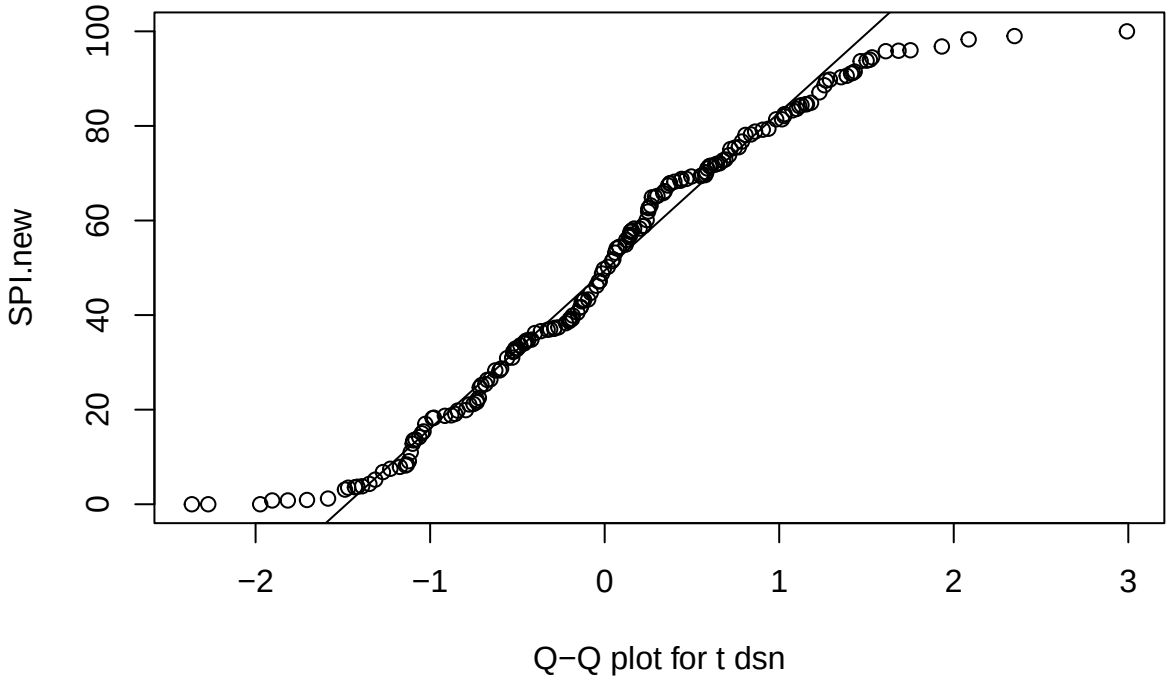

```
qqnorm(SPI.new)
qqline(SPI.new)
```




```
#Make a Q-Q plot against the generating distribution by:
qqplot(rnorm(250), SPI.new, xlab = "Q-Q plot for norm dsn")
qqline(SPI.new)
```




```
qqplot(rt(250, df = 50), SPI.new, xlab = "Q-Q plot for t dsn")
qqline(SPI.new)
```


SPI.new
0
20
40
60
80
100
−2
−1
0
1
2
3
Q−Q plot for t dsn