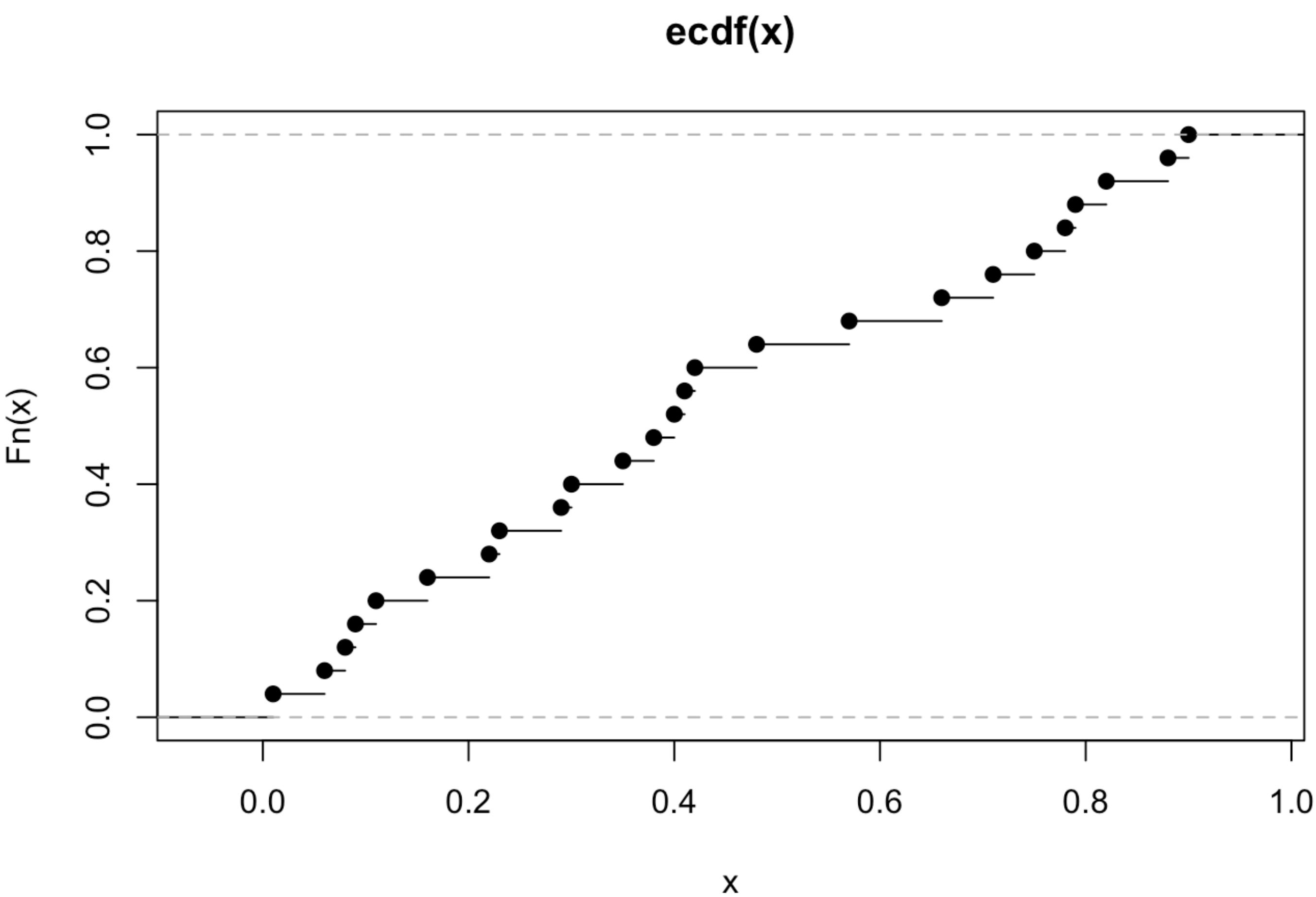
Emprical Distribution

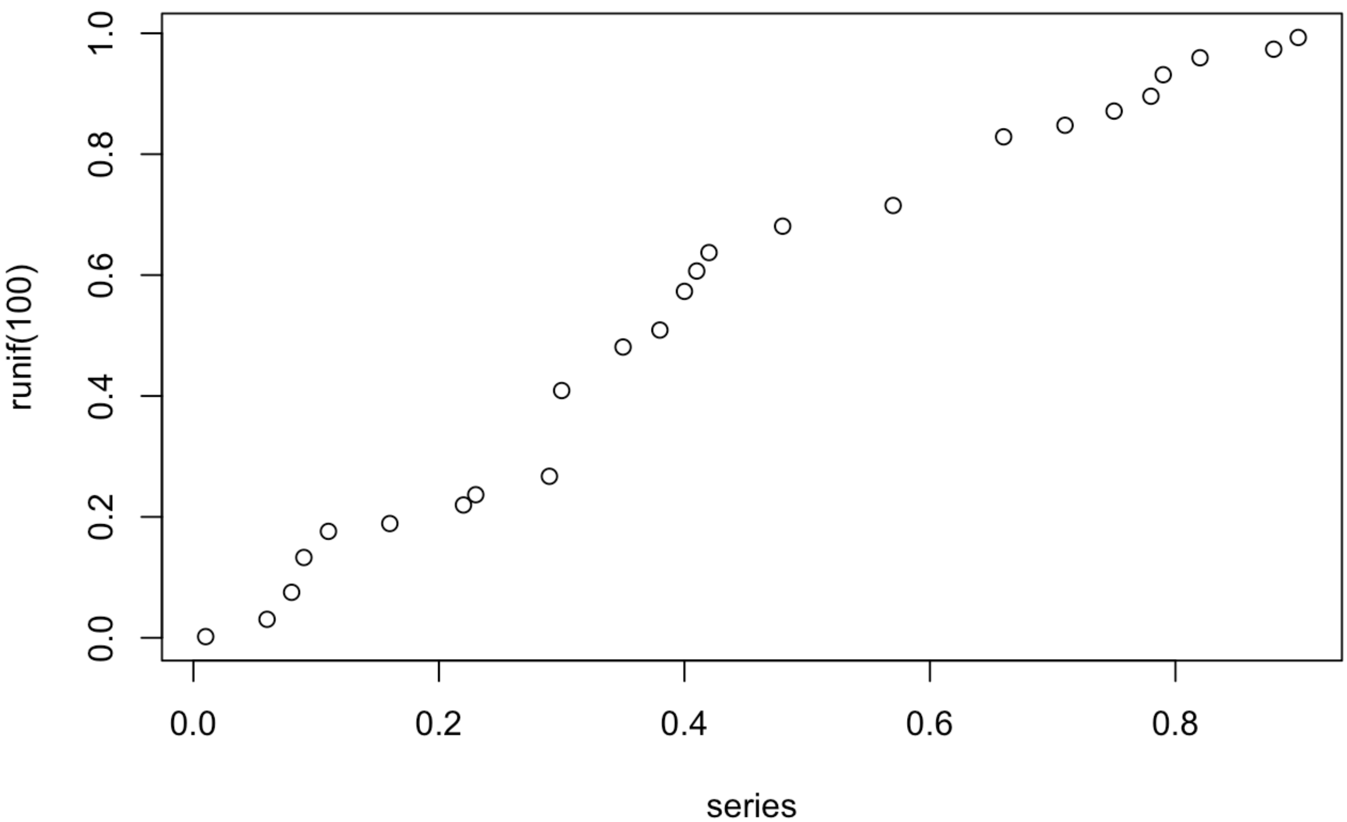
Author: “Diptanshu Singh”

Question 1

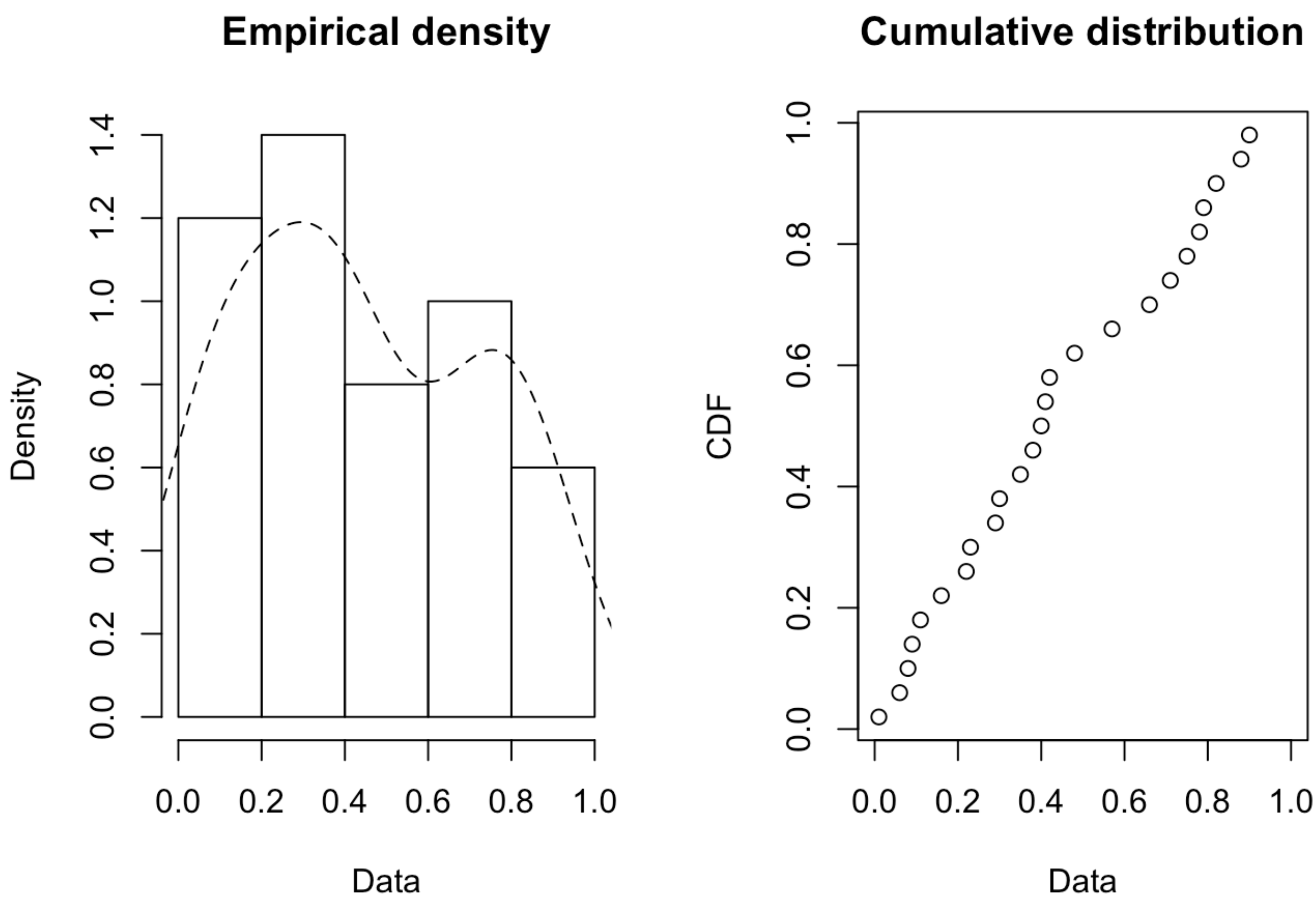
```
tx <- readtext("maybe_uniform.txt")
tx <- strsplit(tx[1,2], c("\n"))[[1]]

series <- c()
for ( i in c(1:length(tx))){
    series <- c(series,as.numeric(strsplit(tx[i], " ")[[1]]))
}

plot.ecdf(series)
```



```
qqplot(x = series, y = runif(100) )
```



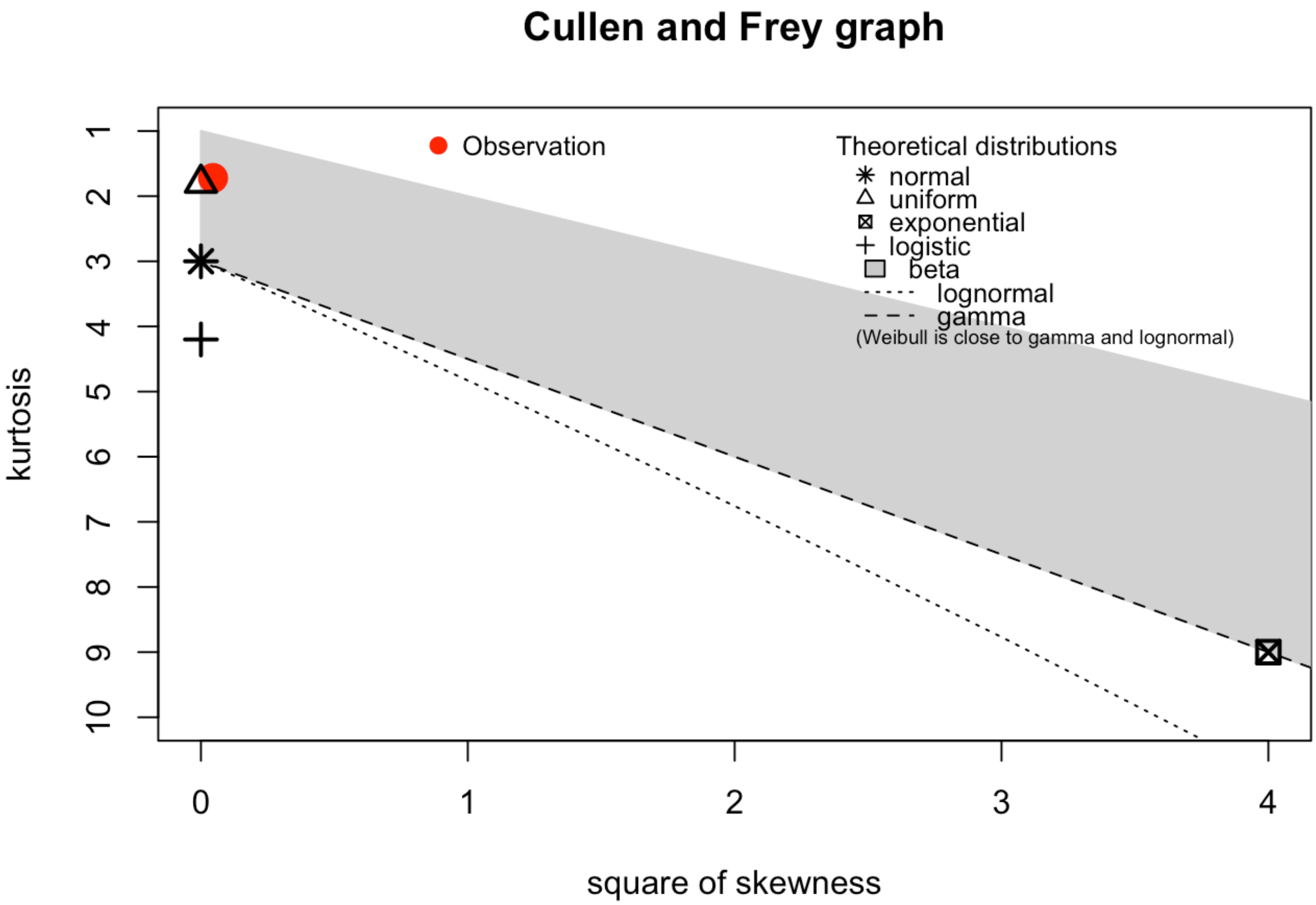

```
# This plot seems like it comes from uniform

#Calculating using the D-statistics:
plotdist(series, demp = TRUE)
```




```
# Probably this distribution is not uniform as visible from the emprical probability
distribution

#Performing KS test
ks.test(series, "punif")
```

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

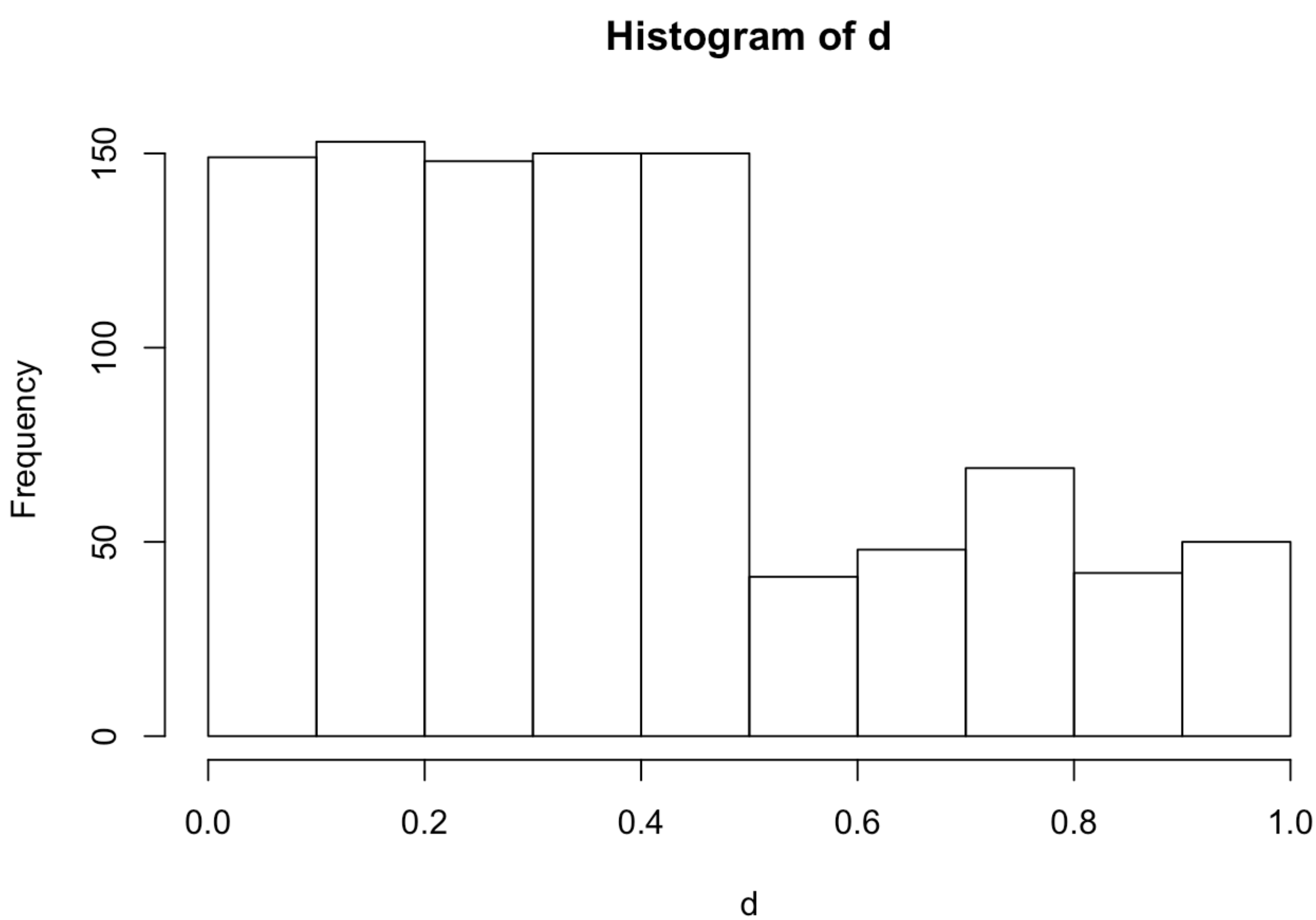
```
# The probability that this distribution came from uniform is 35%
# We will not be surprised if this sample came from uniform distribution

descdist(series, obs.col = "red")
```




```
## 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
```

```
#Also on the cullen and frey graph, this lies very close to the uniform distribution
```

```
# Comparison to other distribution
# Creating a random sample for other distribution

d1 <- runif(750)/2
d2 <- 0.5 + runif(250)/2

d <- c(d1,d2)

hist(d)
```



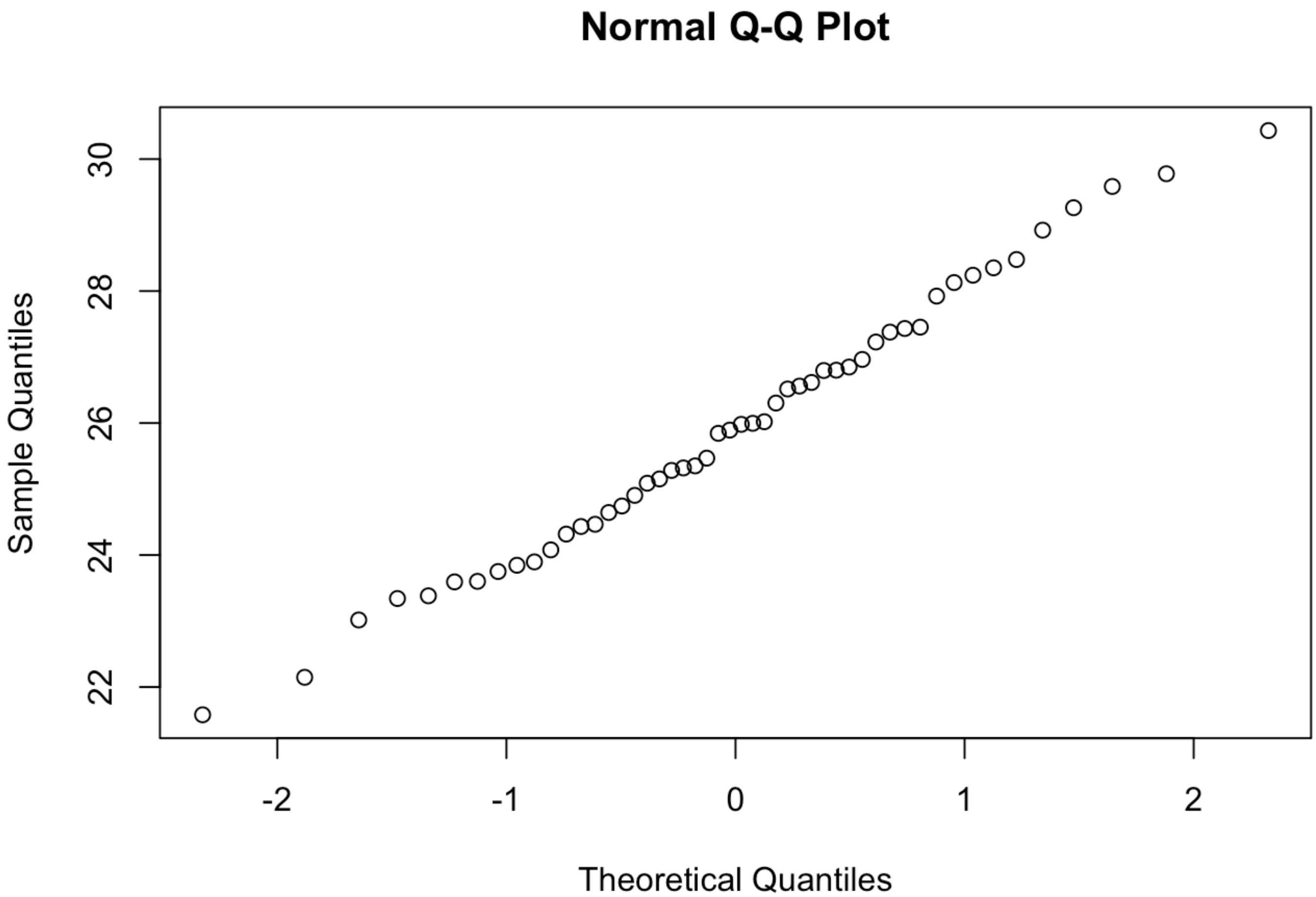

```
# Histogram shows that d follows distribution 3 as given in the question

#Performing KS test
ks.test(series, d)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  series and d
## D = 0.14, p-value = 0.7255
## alternative hypothesis: two-sided
```

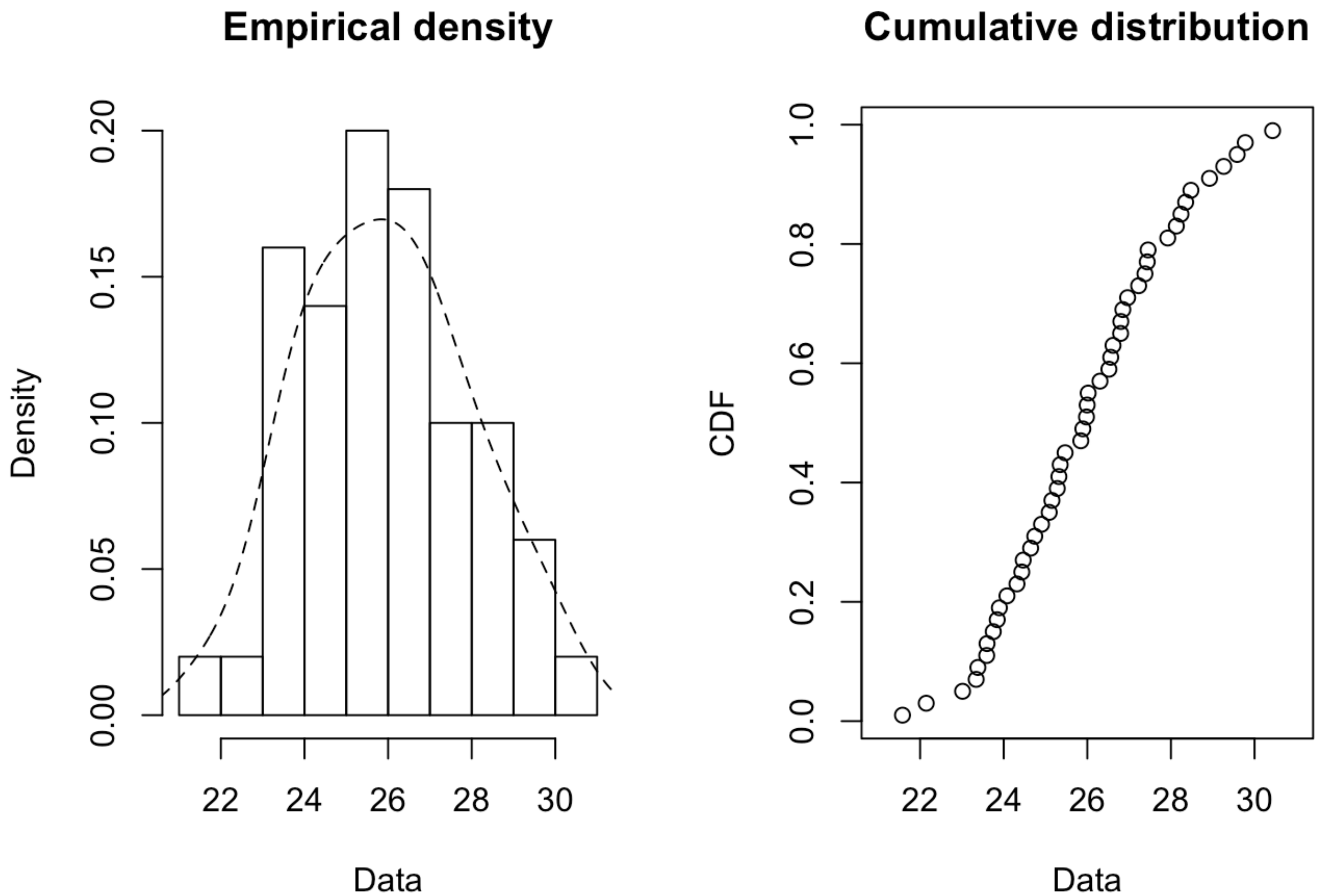
```
# The probability that this distribution came from uniform is 55%
# Its more probable to come to distribution 3, then to come from uniform distribution
# The d statistic is also less compared to this to when we used the uniform distribut
ion
```

Question 2

```
tx <- readtext("maybe_normal.txt")
tx <- strsplit(tx[1,2], c("\n"))[[1]]

series <- c()
for ( i in c(1:length(tx))){
    series <- c(series,as.numeric(strsplit(tx[i], " ")[[1]]))
}


# The QQnorm plot shows that the distribution almost follows a normal distribution
qqnorm(series)
```



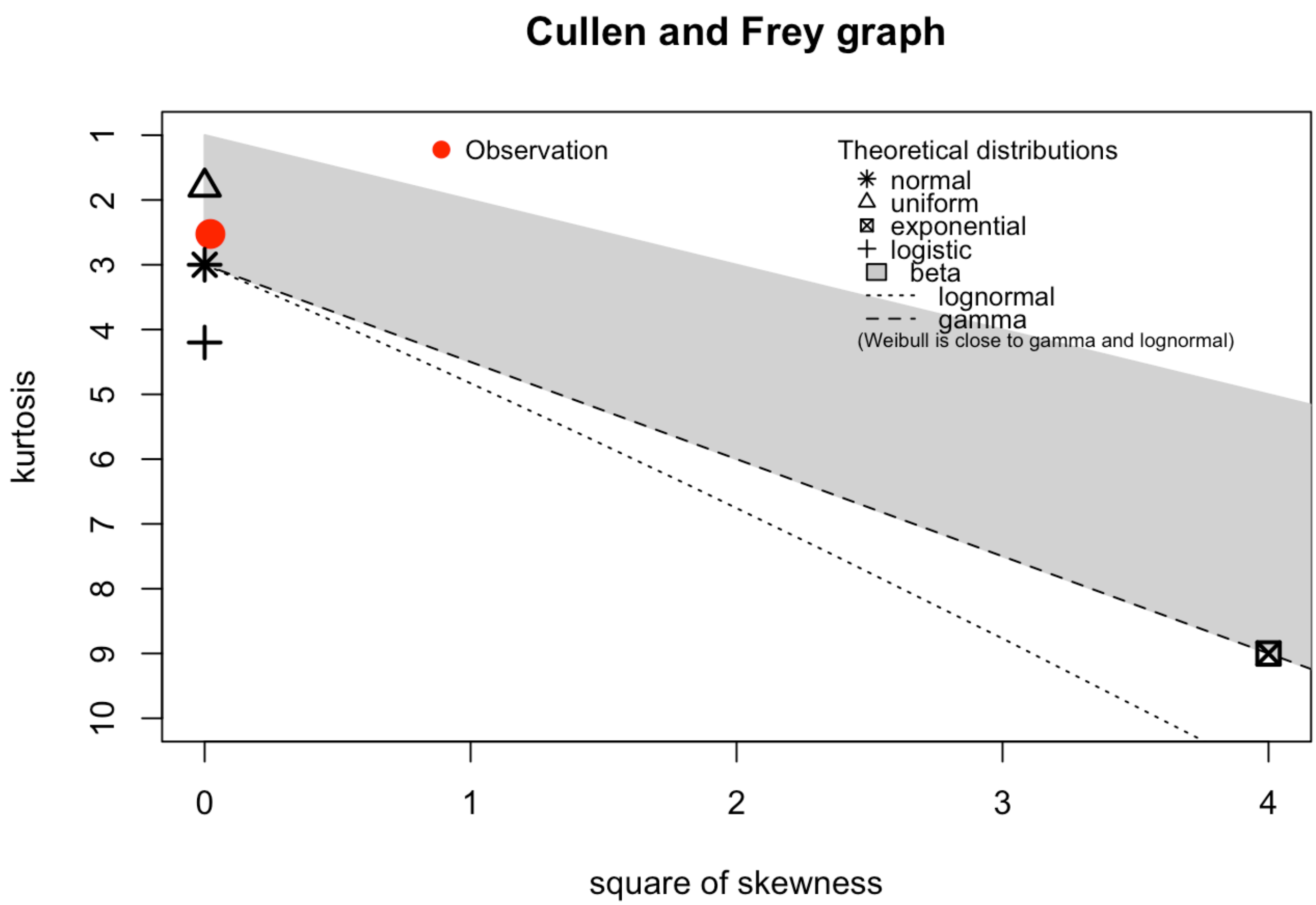

```
# Calculating distance statistics for this distribution
plotdist(series, demp = TRUE)
```




```
# Probably this distribution is normal as visible from the emprical probability distr
ibution

st_series = (series -  mean(series))/ sd(series)
#Performing KS test
ks.test( st_series , "pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  st_series
## D = 0.053959, p-value = 0.997
## alternative hypothesis: two-sided
```

```
# The probability that this distribution came from standard normal is 99.97%

descdist(st_series, obs.col = "red")
```




```
## summary statistics
## ------
## min:  -2.136523   max:  2.198138
## median:  -0.002976928
## mean:  1.004145e-16
## estimated sd:  1
## estimated skewness:  0.1467526
## estimated kurtosis:  2.525013
```

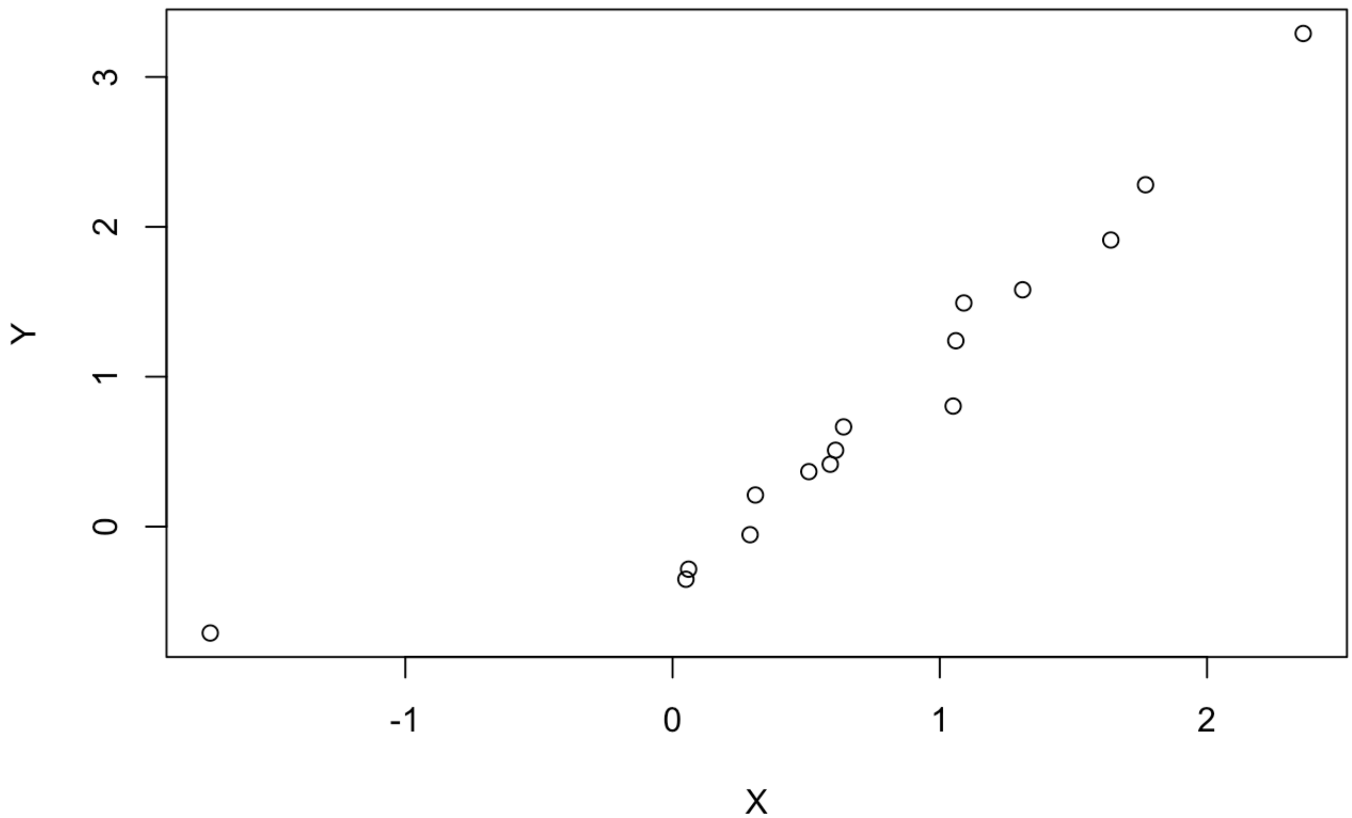
```
#Also on the cullen and frey graph, this lies very close to the normal distribution
```

Question 3

```
#Series 1
tx <- readtext("maybe_same_1.txt")
tx <- str_replace(tx[1,2],"−","-")
tx <- strsplit(tx, c("\n"))[[1]]
series <- c()
for ( i in c(1:length(tx))){
    series <- c(series, as.numeric(strsplit(tx[i], " ")[[1]]))
}
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
X <- series

# Series 2
tx <- readtext("maybe_same_2.txt")
tx <- str_replace_all(tx[1,2],"−","-")
tx <- strsplit(tx, c("\n"))[[1]]
series <- c()
for ( i in c(1:length(tx))){
    series <- c(series, as.numeric(strsplit(tx[i], " ")[[1]]))
}

Y <- series

qqplot(X,Y)
```




```
# This plot shows that the distributions are not very simillar

#Calculating d statistics
ks.test(X,Y)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  X and Y
## D = 0.1875, p-value = 0.866
## alternative hypothesis: two-sided
```

```
#The KS test says that there is 33% chance of the distribution being simillar


ks.test(X+2, Y)
```

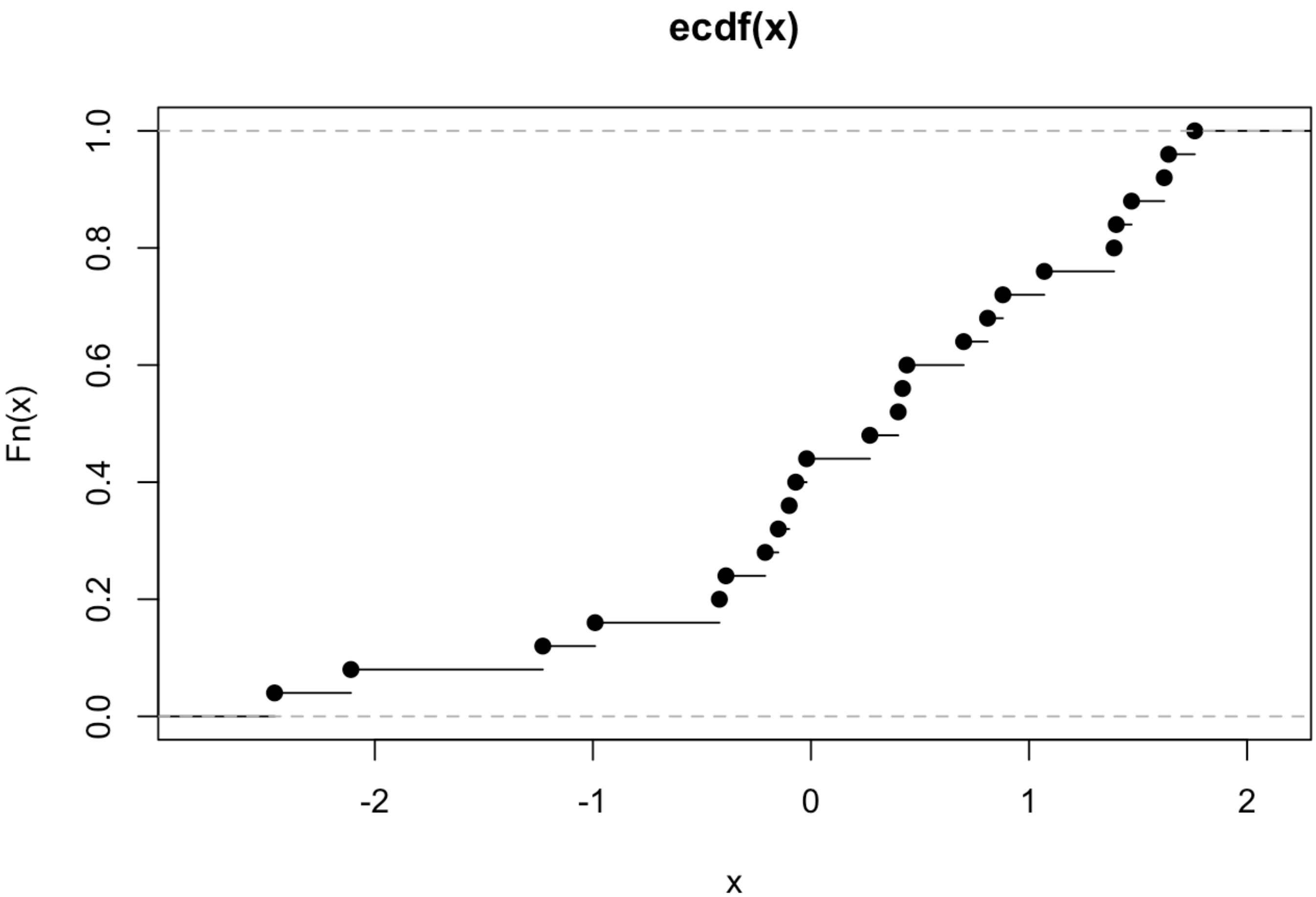
```
## Warning in ks.test(X + 2, Y): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  X + 2 and Y
## D = 0.7875, p-value = 3.258e-05
## alternative hypothesis: two-sided
```

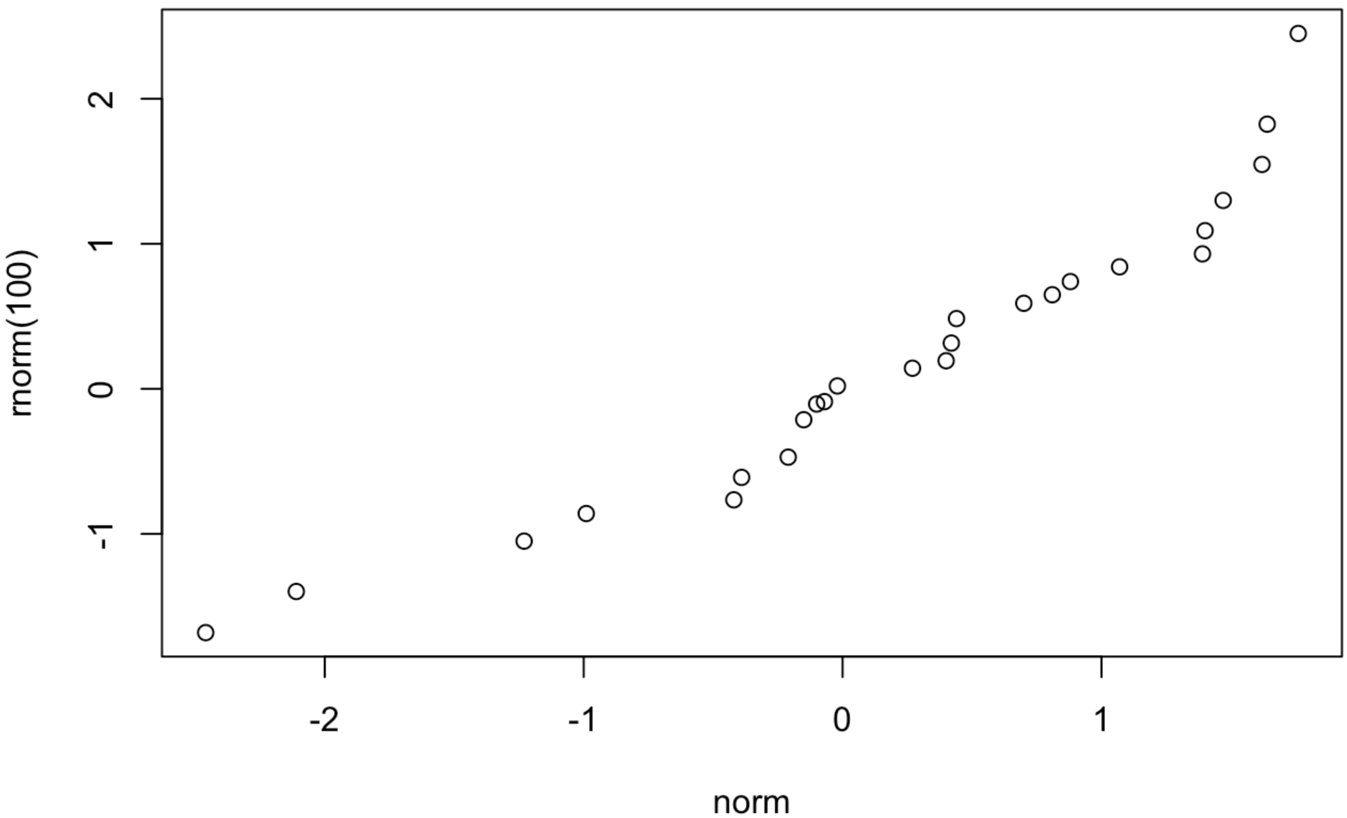
```
# The KS test finds the distribution to be very different
```

Question 4

```
norm <- readRDS("norm_sample.Rdata")

#plot ecdf
plot.ecdf(norm)
```




```
# This plot seems to be very off from the normal distribution
qqplot(norm,rnorm(100))
```




```
#Standardizing variable to find d statistics
norm <- (norm - mean(norm)) / sd(norm)

#KS test
ks.test(norm, rnorm(100))
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  norm and rnorm(100)
## D = 0.17, p-value = 0.5784
## alternative hypothesis: two-sided
```

```
# We cannot reject the hypothesius that they are from different distribution
```

Question 5

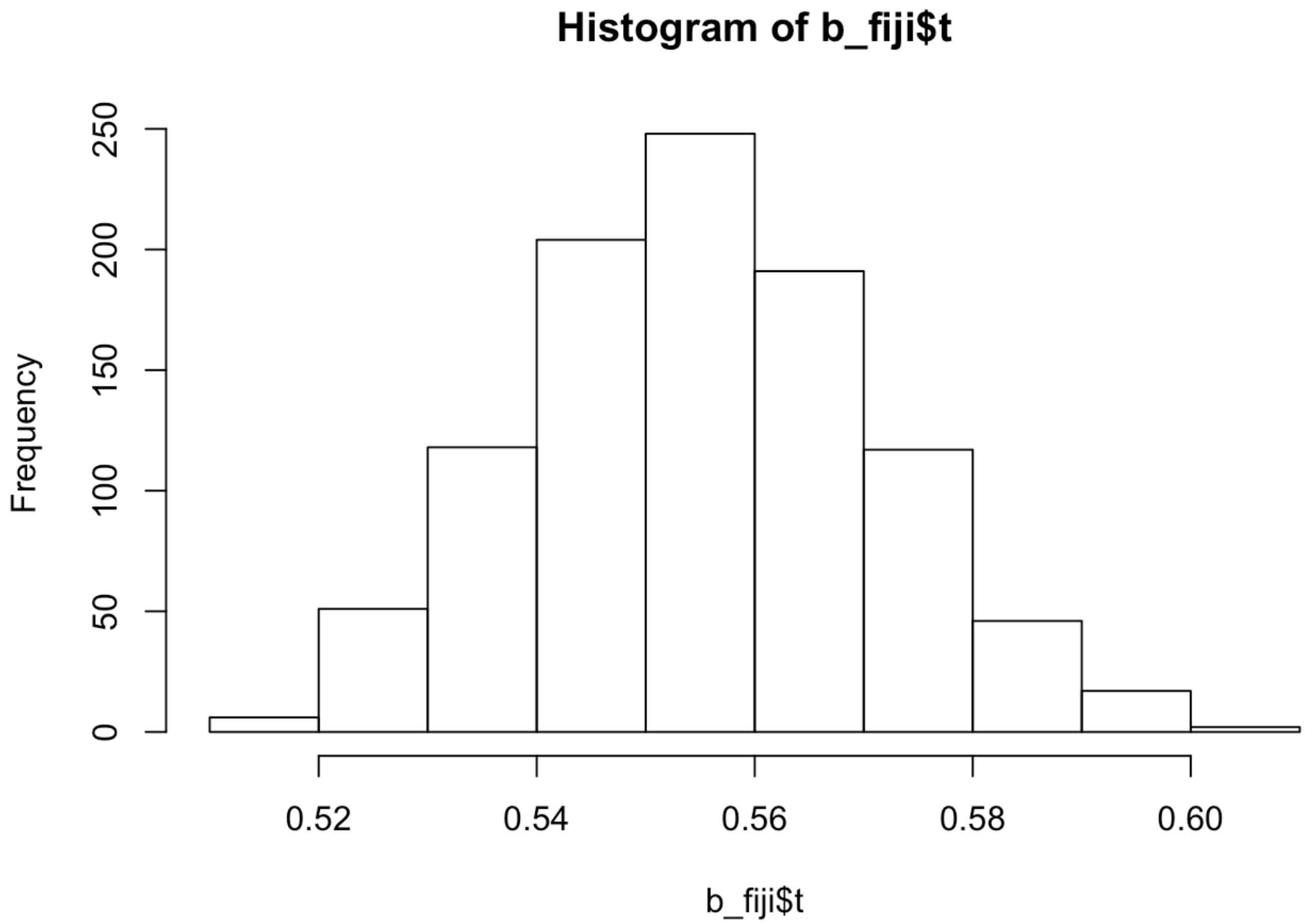
```
fj <- read_table("fijiquakes.dat")
```

```
## Parsed with column specification:
## cols(
##   Obs. = col_double(),
##   lat = col_double(),
##   long = col_double(),
##   depth = col_double(),
##   mag = col_double(),
##   stations = col_double()
## )
```

```
series <- c(fj["mag"])[[1]]

series = sort(series)
cdf_49 <- min(which (series > 4.899999))
cdf_43 <- min(which (series > 4.299999))

val = ( cdf_49 - cdf_43 ) / length(series)
print(val)
```

```
## [1] 0.557
```

```
# Creating the function for calculating the statistics required
rsq <- function( data = series, indices){
  series = series[indices]
  series = sort(series)
  cdf_49 <- min(which (series > 4.899999))
  cdf_43 <- min(which (series > 4.299999))
  val = ( cdf_49 - cdf_43 ) / length(series)
  return(val)
}

# Creating bootstrapped samples for calculating confidence interval
b_fiji <- boot(series, statistic = rsq , 1000 )

hist(b_fiji$t)
```




```
print("The mean value is ")
```

```
## [1] "The mean value is "
```

```
print(mean(b_fiji$t))
```

```
## [1] 0.55578
```

```
# 95% confidence interval
boot.ci(b_fiji, type="bca")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = b_fiji, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.527,  0.591 )
## Calculations and Intervals on Original Scale
```

```
mu <- function(series, indices ){
  d <- series[indices]
  return (mean(d))
}

fj <- read_table("faithful.dat")
```

```
## Parsed with column specification:
## cols(
##   `Old Faithful Geyser Data` = col_character()
## )
```

```
tab <- strsplit(fj$`Old Faithful Geyser Data`[14:285], "\t")%>% unlist

series = c()
for ( i in 1:length(tab)){
  series = c(series,(as.numeric( substr(tab[i] , 9 ,13))))
}

med <- function( data = series, indices){
  series = series[indices]
  return(median(series))
}

b_faith <- boot(series, statistic = med , 1000 )

# 90 percentile confidence interval
print("90% confidence interval for median of waiting time")
```

```
## [1] "90% confidence interval for median of waiting time"
```

```
quantile(b_faith$t,c(0.05,0.95))
```

```
##     5%    95%
## 3.8415 4.1000
```