Assignment 2

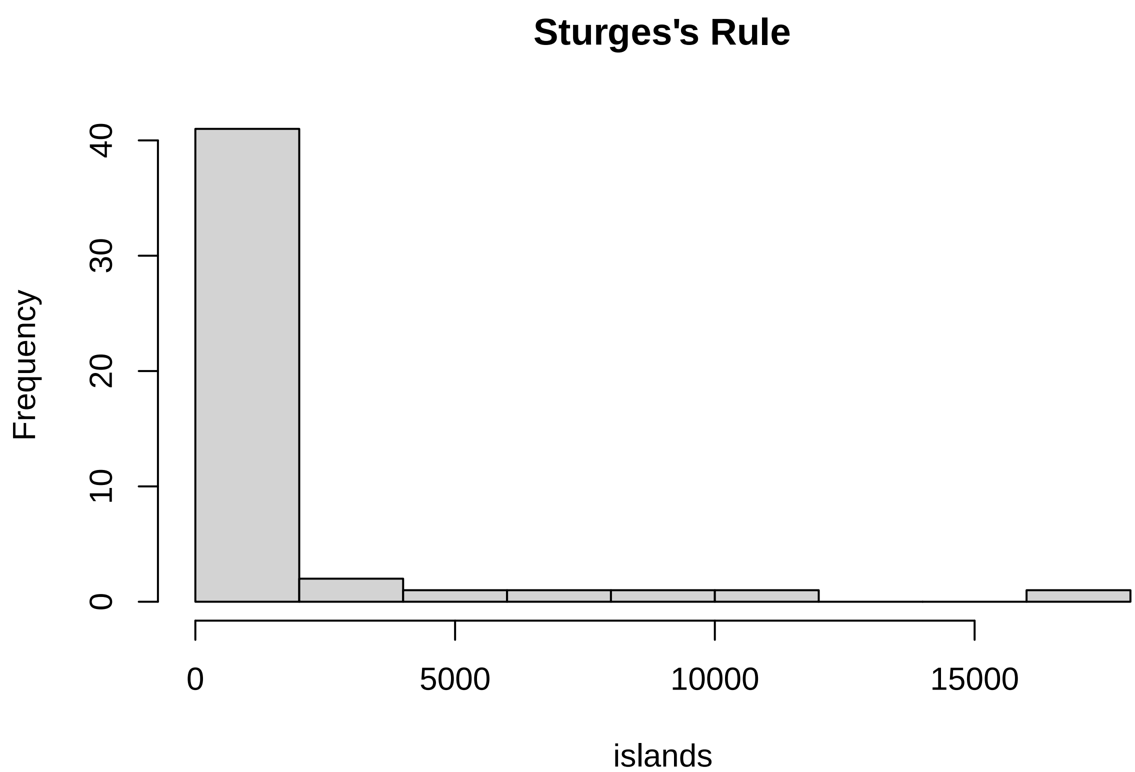
Problem 1

Construct histograms for the islands vector (it is a built-in vector in R, simply type "islands") using breaks based on Sturges's and Scott's rules. Which one looks more informative to you?

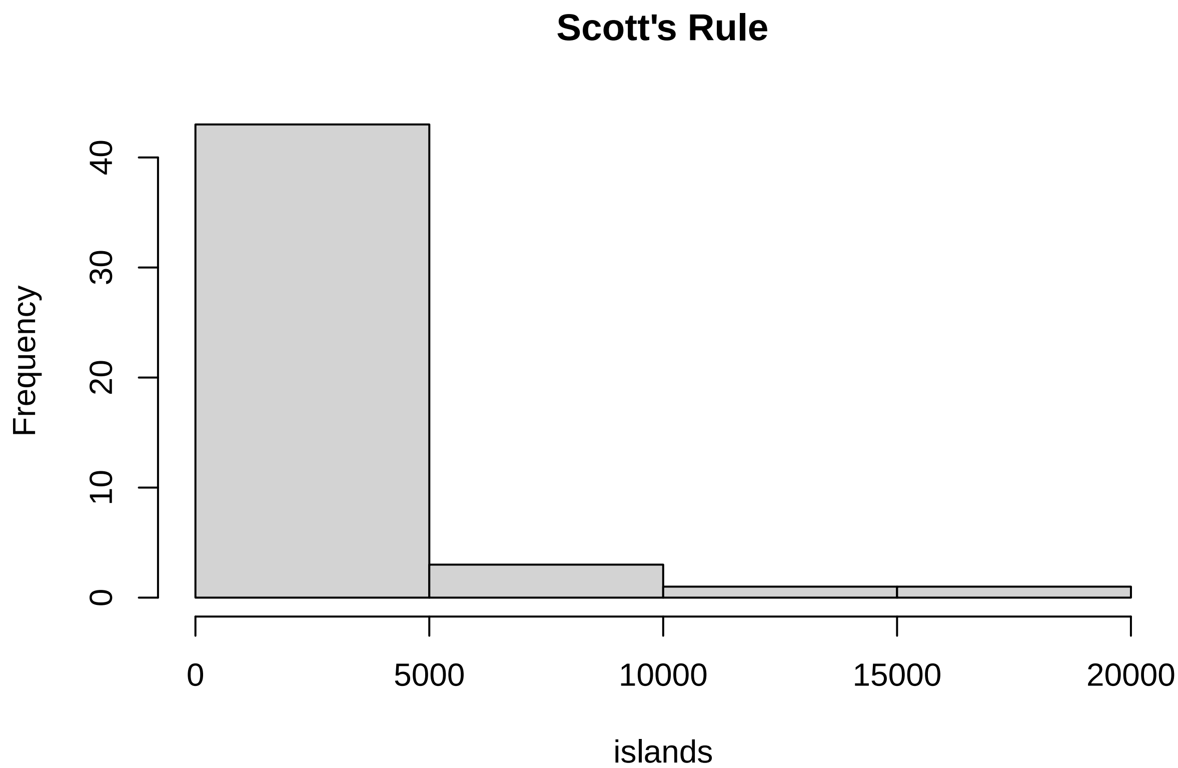
```
islands
```

```
##           Africa        Antarctica              Asia         Australia
##            11506              5500             16988              2968
##     Axel Heiberg            Baffin             Banks            Borneo
##               16               184                23               280
##          Britain           Celebes             Celon              Cuba
##               84                73                25                43
##            Devon         Ellesmere            Europe         Greenland
##               21                82              3745               840
##           Hainan        Hispaniola          Hokkaido            Honshu
##               13                30                30                89
##          Iceland           Ireland              Java            Kyushu
##               40                33                49                14
##            Luzon        Madagascar          Melville          Mindanao
##               42               227                16                36
##         Moluccas       New Britain        New Guinea  New Zealand (N)
##               29                15               306                44
##  New Zealand (S)      Newfoundland     North America     Novaya Zemlya
##               58                43              9390                32
##  Prince of Wales          Sakhalin     South America       Southampton
##               13                29              6795                16
##      Spitsbergen           Sumatra            Taiwan          Tasmania
##               15               183                14                26
## Tierra del Fuego             Timor         Vancouver          Victoria
##               19                13                12                82
```

```
hist(islands, breaks = "Sturges", main = "Sturges's Rule")
```




```
hist(islands, breaks = "Scott", main = "Scott's Rule")
```

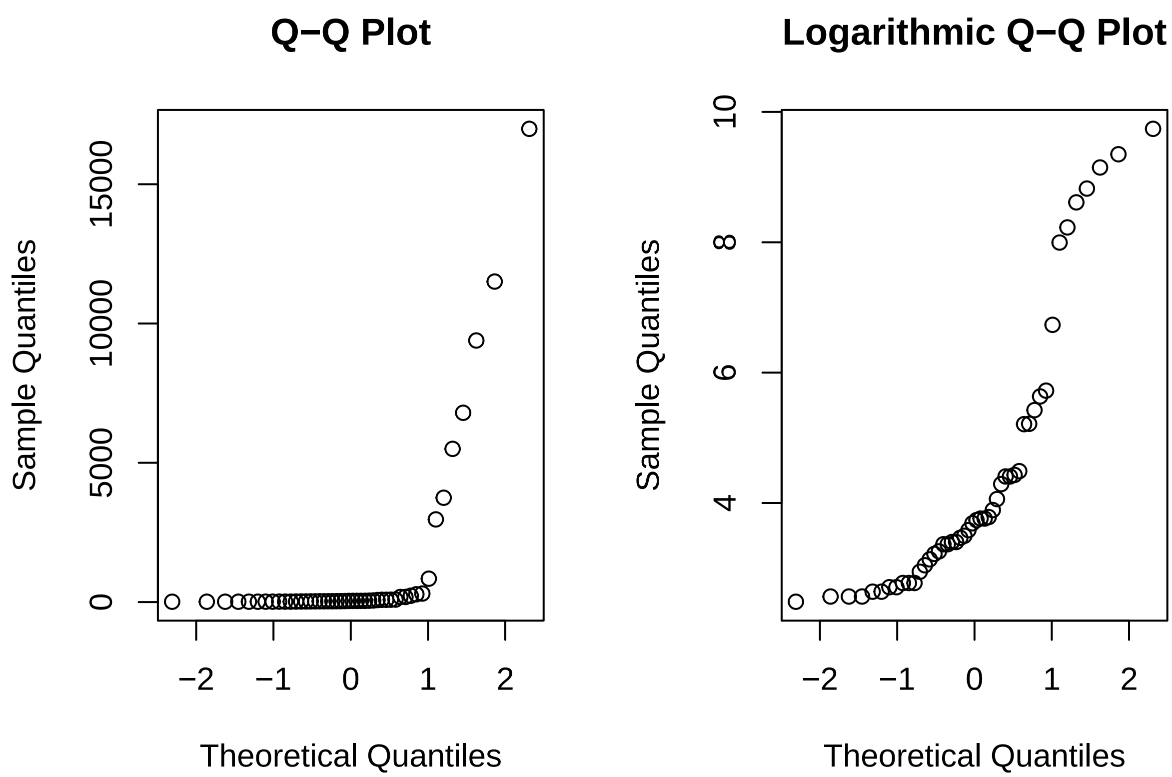



```r
# Even though Scott's rule has more bins overall, Sturges's rule creates two bars per bin
# Especially in the 0-5000 bin which makes most of the dataset, Sturge's rule is clearer
# With Scott's rule, the dataset looks even more skewed to the right, which is accurate
```

Problem 2

Construct a Q-Q plot for both islands and its log against a normal distribution. Which follows a normal distribution more closely?

```r
par(mfrow = c(1, 2)) # 1x2 layout makes it better for direct comparison
qqnorm(islands, main = "Q-Q Plot")
qqnorm(log(islands), main = "Logarithmic Q-Q Plot")
```




```
# neither truly follows a normal distribution upon inspection
# but logarithmic plot resembles more the bell-shaped normal distribution curve
```

Problem 3

Generate 1000 Uniform(0, 1) pseudorandom variables using the runif() function, assigning them to a vector called U. Use the seed 09062024.

```
set.seed(09062024) # for future reference
U <- runif(1000)

sample_mean <- mean(U)
sample_variance <- var(U)
sample_sd <- sd(U)
sample_mean
```

```
## [1] 0.5079104
```

```
sample_variance
```

```
## [1] 0.07865998
```

```r
sample_sd
```

```
## [1] 0.2804639
```

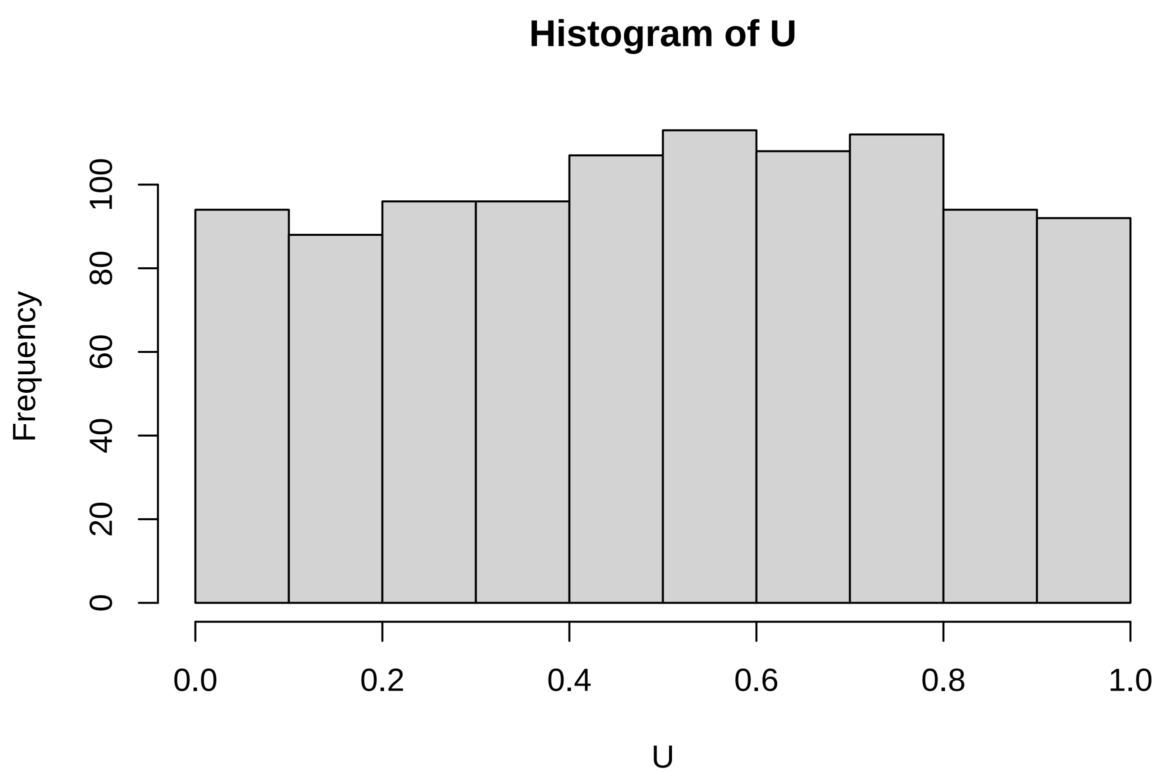
```r
# For Uniform(0,1), the theoretical mean is 0.5
# Theoretical variance is (1-0)^2 / (12) = 1 / 12 = 0.083333...
# Naturally, it follows that theoretical sd would be sqrt(12) = 0.2887
# Sample mean is greater, sample variance is smaller, and sample sd is smaller

proportion_less_than <- mean(U < 0.6)
proportion_less_than
```

```
## [1] 0.594
```

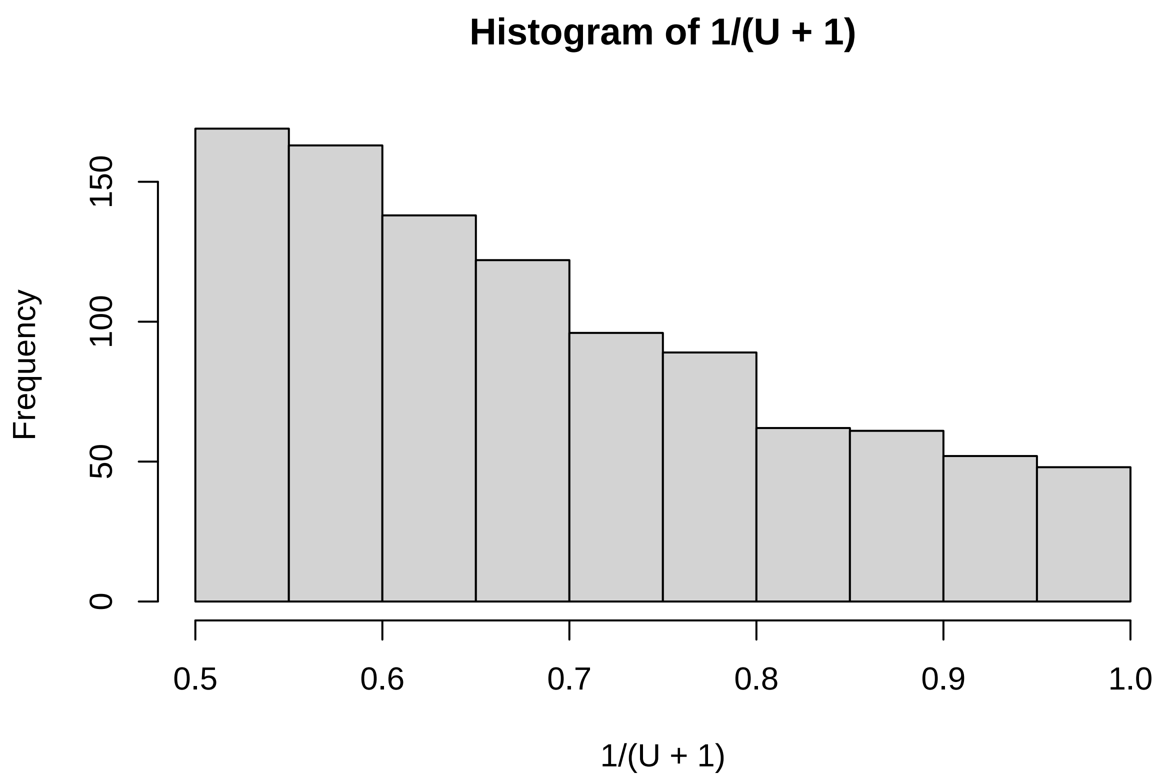
```r
# theoretical probability that a Uniform(0, 1) random variable < 0.6 is 0.6
# so the theoretical probability is greater

expected_value <- mean(1 / (U + 1)) # EV is just another way of saying the mean
expected_value
```

```
## [1] 0.6879805
```

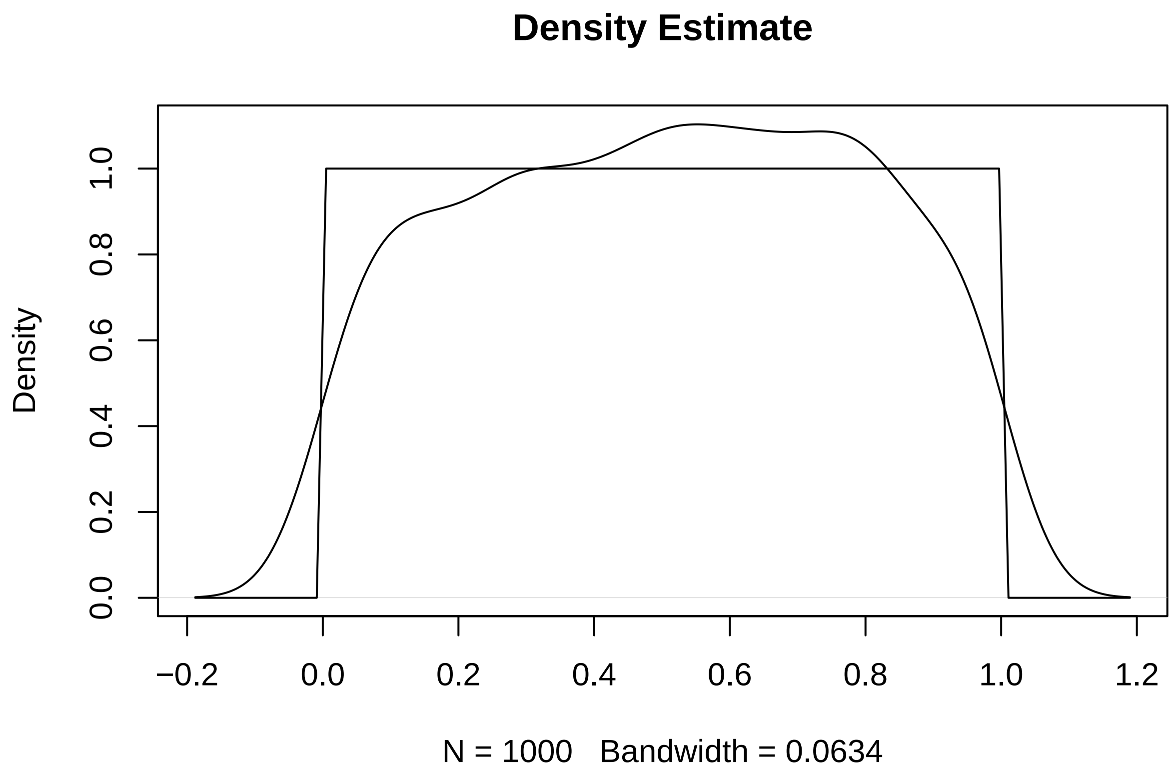
```r
hist(U, main = "Histogram of U", xlab = "U")
```



```
hist(1 / (U + 1), main = "Histogram of 1/(U + 1)", xlab = "1/(U + 1)")
```




```
par(mfrow = c(1, 1)) # change back settings to original

plot(density(U), main = "Density Estimate")
curve(dunif(x, min = 0, max = 1), add = TRUE)
```



Problem 4

Let X be a Binomial(20, 0.3) random variable. Use built-in R functions (not a math formula) to find the exact numerical values of the following quantities:

```r
p_less <- pbinom(5, size = 20, prob = 0.3)
p_less
```

```
## [1] 0.4163708
```

```r
p_equals <- dbinom(5, size = 20, prob = 0.3)
p_equals
```

```
## [1] 0.1788631
```

```r
p_range <- pbinom(7, size = 20, prob = 0.3) - pbinom(4, size = 20, prob = 0.3)
p_range
```

```
## [1] 0.534764
```

```r
quantile_90 <- qbinom(0.9, size = 20, prob = 0.3)
quantile_90
```

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

Problem 5

Generate 30 Binomial(20, 0.3) random variables. (You can use any random seed.)

```
set.seed(123)
data <- rbinom(30, size = 20, prob = 0.3)
data_quantile_90 <- quantile(data, 0.9)
data_quantile_90
```

```
## 90%
## 9.1
```

```
quantile_90
```

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

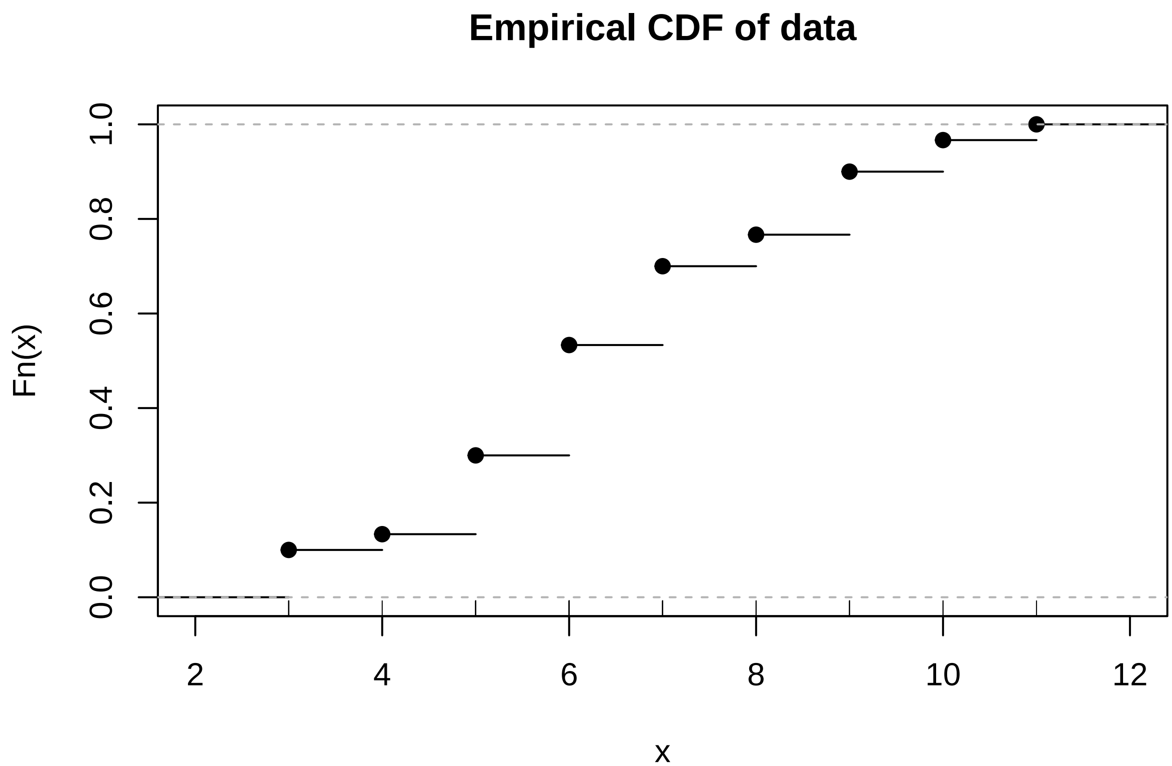
```
# so sample quantile from data is slightly greater

prop_less_than_equal_5 <- mean(data <= 5)
prop_less_than_equal_5
```

```
## [1] 0.3
```

```
p_less
```

```
## [1] 0.4163708
```

```
# theoretical proportion is greater

plot(ecdf(data), main = "Empirical CDF of data")
rug(data)
```


Empirical CDF of data
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
x