HW3

Xinyue (Ivory) Liu

9/25/2018

25. X is a random variable with a Beta distribution with parameters (2.5, 1.8). Use a suitable simulation to approximate the expected value of log(X) with an error of less than 10^{-3} and explain why you think you have achieved this accuracy. Do not use more than 100 times simulations.

```
beta_r <- log(rbeta(10000000, 2.5, 1.8))
round(mean(beta_r),3)
```

```
## [1] -0.635
```

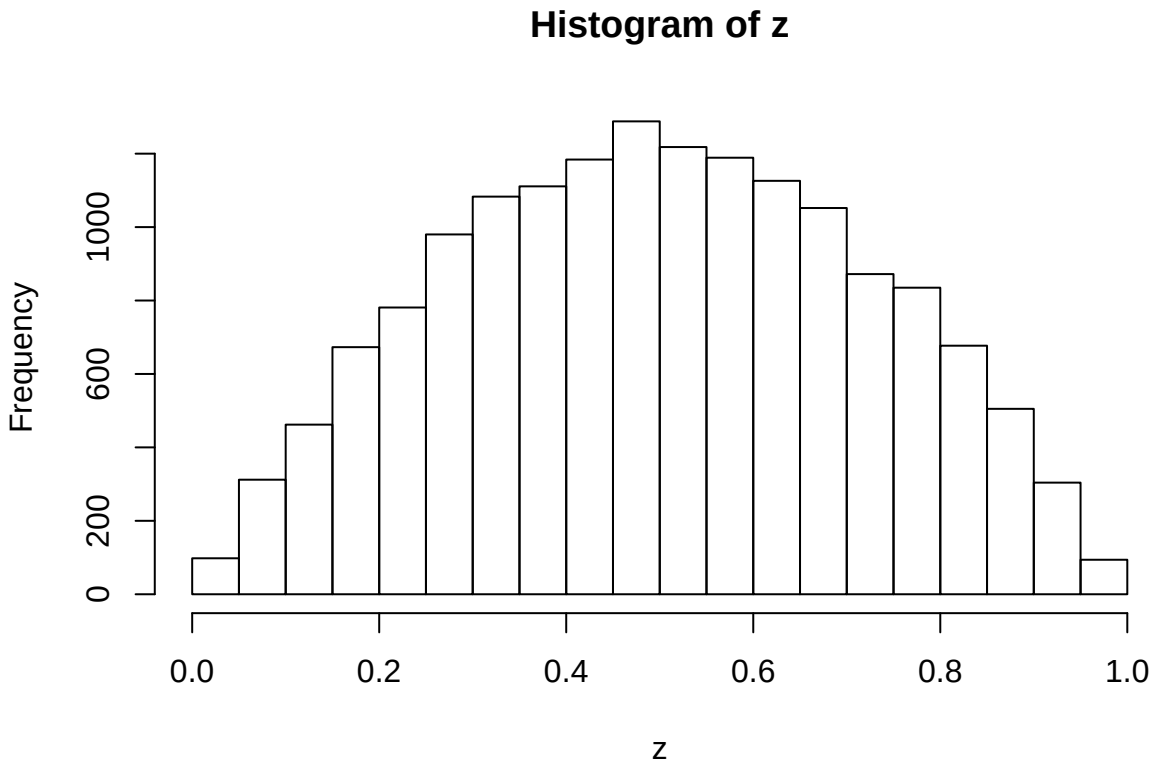
26. X is a random variable with a Beta distribution with parameters (0.8, 1.5). Use a suitable simulation to demonstrate harmonic mean $H(X) = \frac{1}{E(\frac{1}{X})}$ does not exist.

```
s <- seq(1000, 50000, by=1000)
x <- c()
for(n in s){
  beta_r <- 1/(rbeta(n, 0.8, 1.5))
  beta_e <- mean(beta_r)
  beta_h <- 1/beta_e
  x <- append(x, beta_h)
}
summary(x)
```

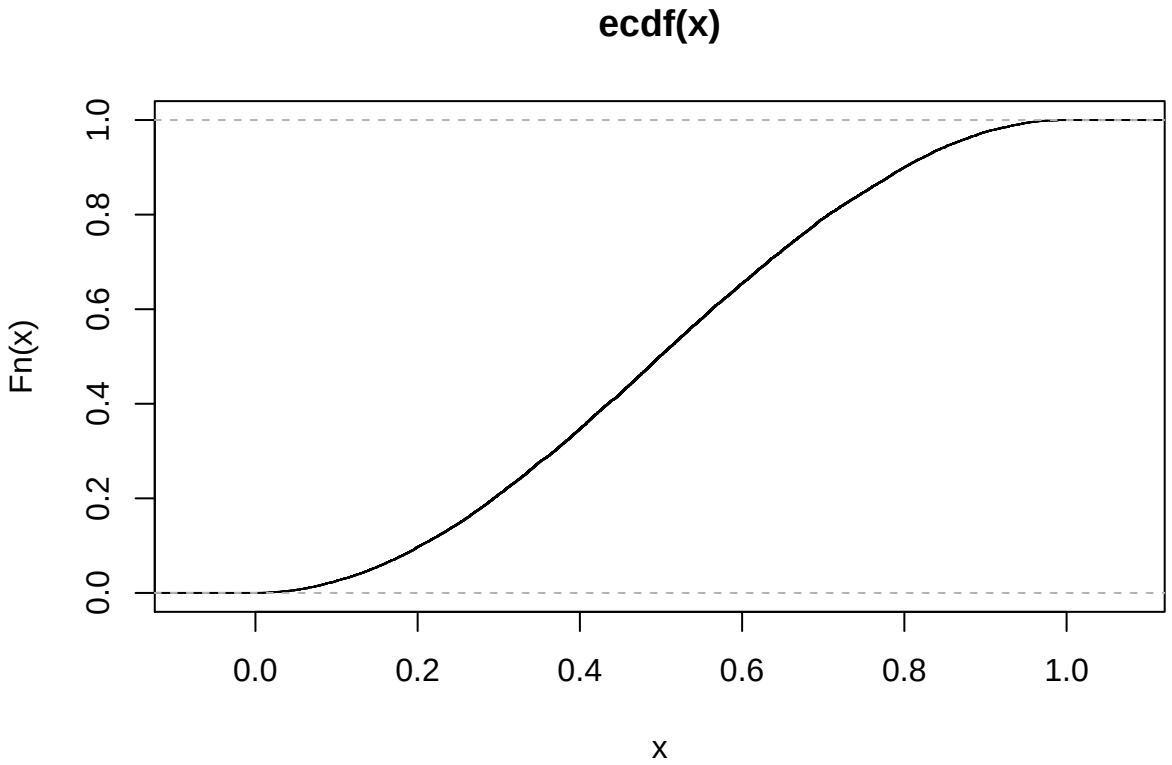
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0005307 0.0041383 0.0069885 0.0080705 0.0124324 0.0195996
```

27. Suppose X and Y are independent random variables that both have uniform U(0,1) distributions. Consider the events $A = X \leq \frac{1}{3}$, $B = Y < sin\pi X$. Then $P(A) = \frac{1}{3}$. Use R and simulations to estimate P(B), P(A|B), P(B|A)

```
x <- runif(25000)
y <- runif(25000)
z <- x[y<sin(pi*x)]
hist(z)
```




```
plot.ecdf(z)
```



$P(A) = P(X \leq \frac{1}{3}) = \frac{1}{3}$

$P(B) = P(Y < sin(\pi X)) = sin(\pi X)$

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(X \leq \frac{1}{3} \cap Y < sin(\pi X))}{P(Y < sin(\pi X))} = \frac{P(X \leq \frac{1}{3} \cap Y < sin(\frac{1}{3}\pi))}{P(Y < sin(\pi X))} = \frac{1}{3} * \frac{\sqrt{3}}{2} / 1 = \frac{\sqrt{3}}{6}$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(X \leq \frac{1}{3} \cap Y < sin(\pi X))}{P(X \leq \frac{1}{3})} = \frac{P(X \leq \frac{1}{3} \cap Y < sin(\frac{1}{3}\pi))}{P(X \leq \frac{1}{3})} = \frac{1}{3} * \frac{\sqrt{3}}{2} / \frac{1}{3} = \frac{\sqrt{3}}{2}$$

28. Breathalyzer can identify a drunk driver with probability 0.98 and falsely identify a sober driver as drunk with probability 0.01.

a) What is the sensitivity of this breathalyzer? What is specificity?

```
p_fail_drunk <- 0.98

p_fail_sober <- 0.01

# therefore:
p_pass_drunk <- 1 - 0.98

p_pass_sober <- 1 - 0.01

# Sensitivity = true positives / (true positives + false negatives)

# Specificity = true negatives / (true negatives + false positives)

sensitivity <- 0.98 / (0.98 + 0.99)
sensitivity
```

```
## [1] 0.4974619
```

```
specificity <- 0.01 / (0.01 + 0.02)
specificity
```

```
## [1] 0.3333333
```

b) About 1 in 500 drivers are drunk. A car is stopped by the police and the driver fails the breathalyzer test. What is the probability that she is drunk?

```
p_drunk <- 1/500
p_fail <- p_fail_drunk * p_drunk + p_fail_sober * (1-p_drunk)
p_drunk_fail <- p_fail_drunk * p_drunk / p_fail
p_drunk_fail
```

```
## [1] 0.1641541
```

30.

```
x <- rgamma(1000, 2.5)
k <- 2
z1 <- x[x<=k]
z2 <- min(x, k)
mean(z1)
```

```
## [1] 1.203062
```

```
mean(z2)
```

```
## [1] 0.03344524
```

31. Consider the baby names data for 2002. Write a function that computes the conditional probability P(gender=F|name=XXX) for a given character string XXX and use it to compute these conditional probabilities for the 10 most common female baby names of that year. For which female baby name of the top 10 is this conditional probability maximal? What does this mean?

```
yob2002 <- read.csv("yob2002.txt",
                    header=FALSE, stringsAsFactors=FALSE)
names(yob2002) <- c("name","gender","count")
first10 <- head(yob2002,10)$name

for(n in first10){
  both <- yob2002[yob2002$name == n,]
  female <- both[both$gender == "F",]
  print(n)
  print(sum(female$count)/sum(both$count)) # P(female | Name)
}
```

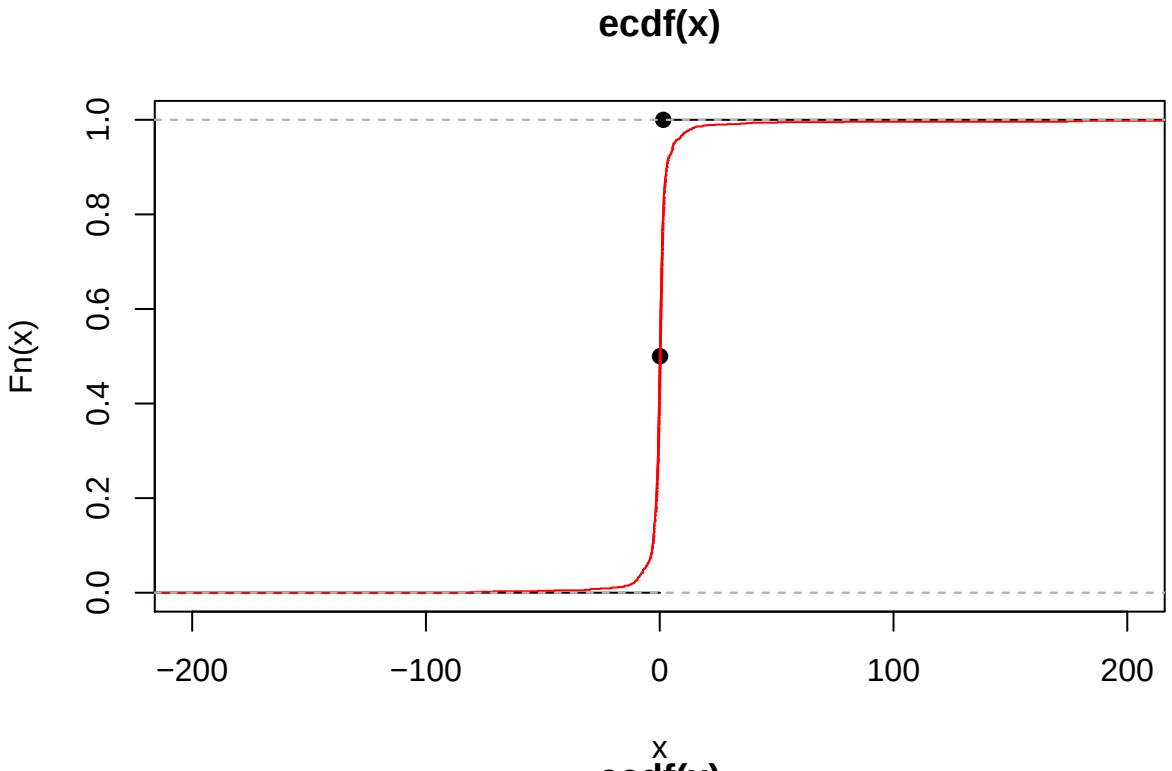
```
## [1] "Emily"
## [1] 0.9986117
## [1] "Madison"
## [1] 0.9952459
## [1] "Hannah"
## [1] 0.99862
## [1] "Emma"
## [1] 0.9990934
## [1] "Alexis"
## [1] 0.8679214
## [1] "Ashley"
## [1] 0.9960384
## [1] "Abigail"
## [1] 0.9986939
## [1] "Sarah"
## [1] 0.9985108
## [1] "Samantha"
## [1] 0.9984333
## [1] "Olivia"
## [1] 0.9988392
```

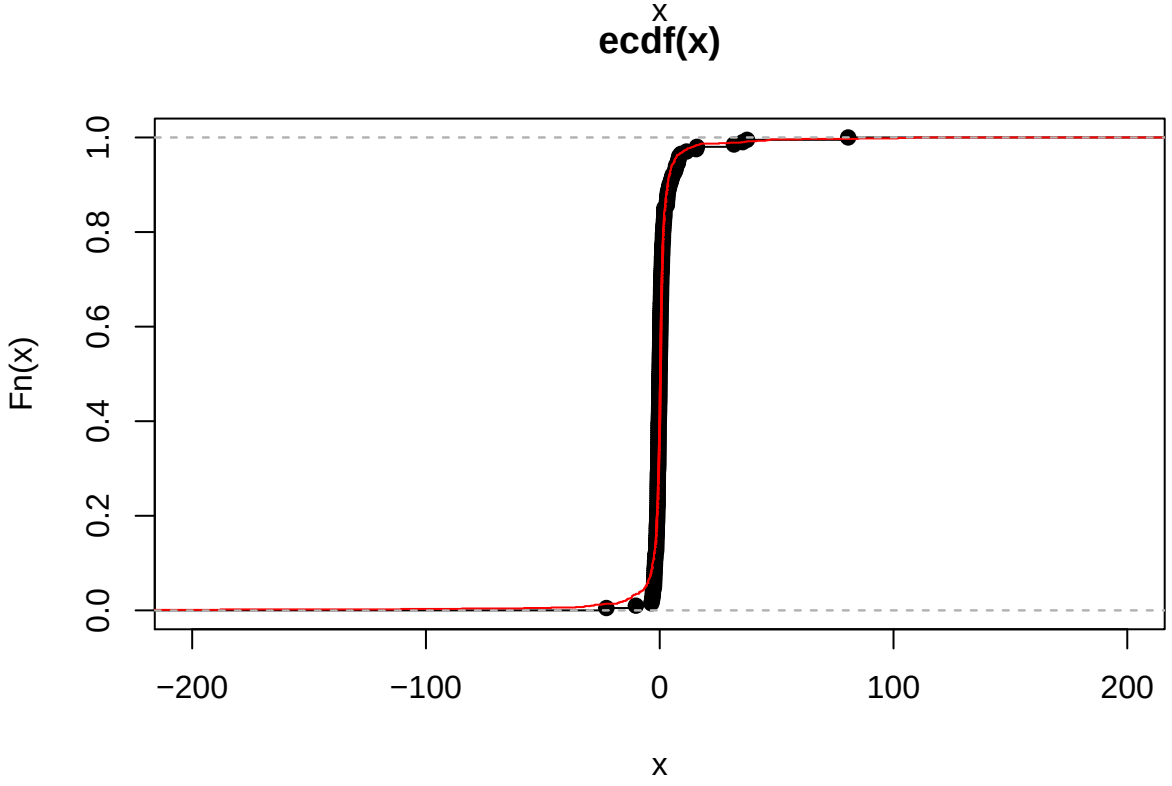
The name "Emma" has the highest conditional probability. This means in year 2002, among the 10 most common female baby names, people with name Emma has a larger probability to be a female than other 9 names. It is the most "feminine" name among the 10.

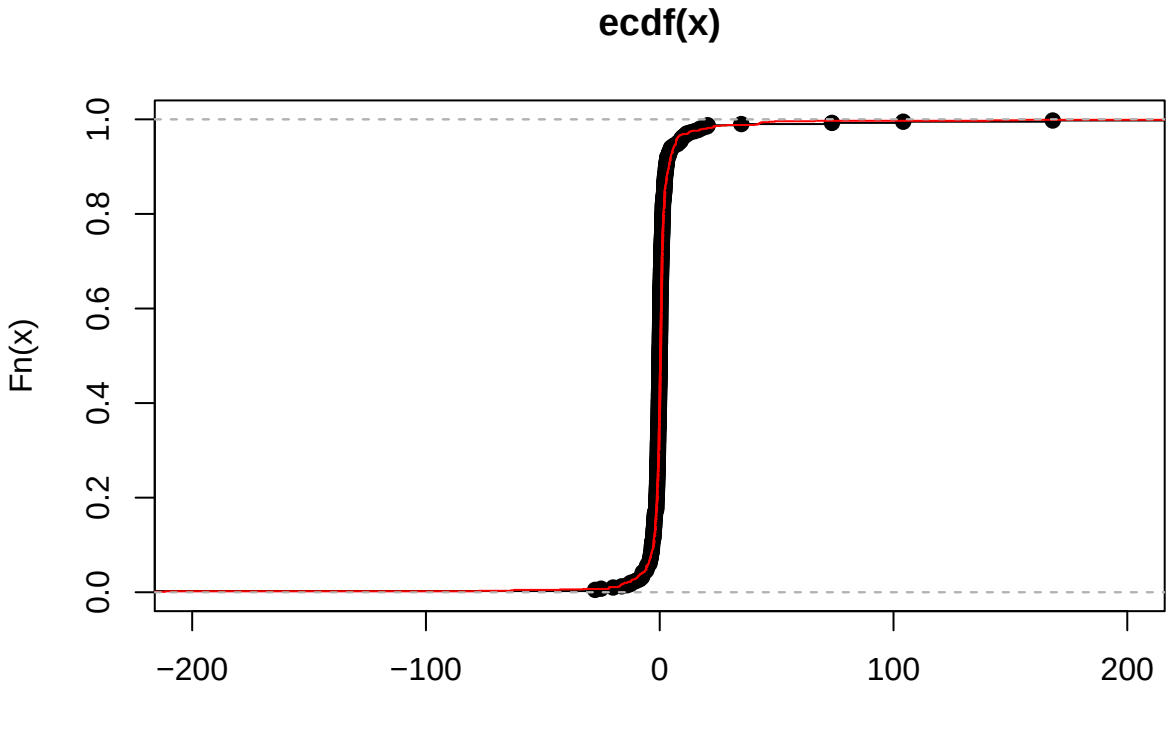
33. The Cauchy distribution has the property that the mean $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ of n independent copies $X_1, ..., X_n$ has again a Cauchy distribution. That's why the sample means do not "settle" on any value - they remain randomly distributed. You can show this by simulating many sample means y_i from simulated Cauchy data with fixed sample size n and making an empirical cdf of the $F(y_i)$. Do this for a range of n values from n=2 to n=1000 and explain why the results demonstrate this property.

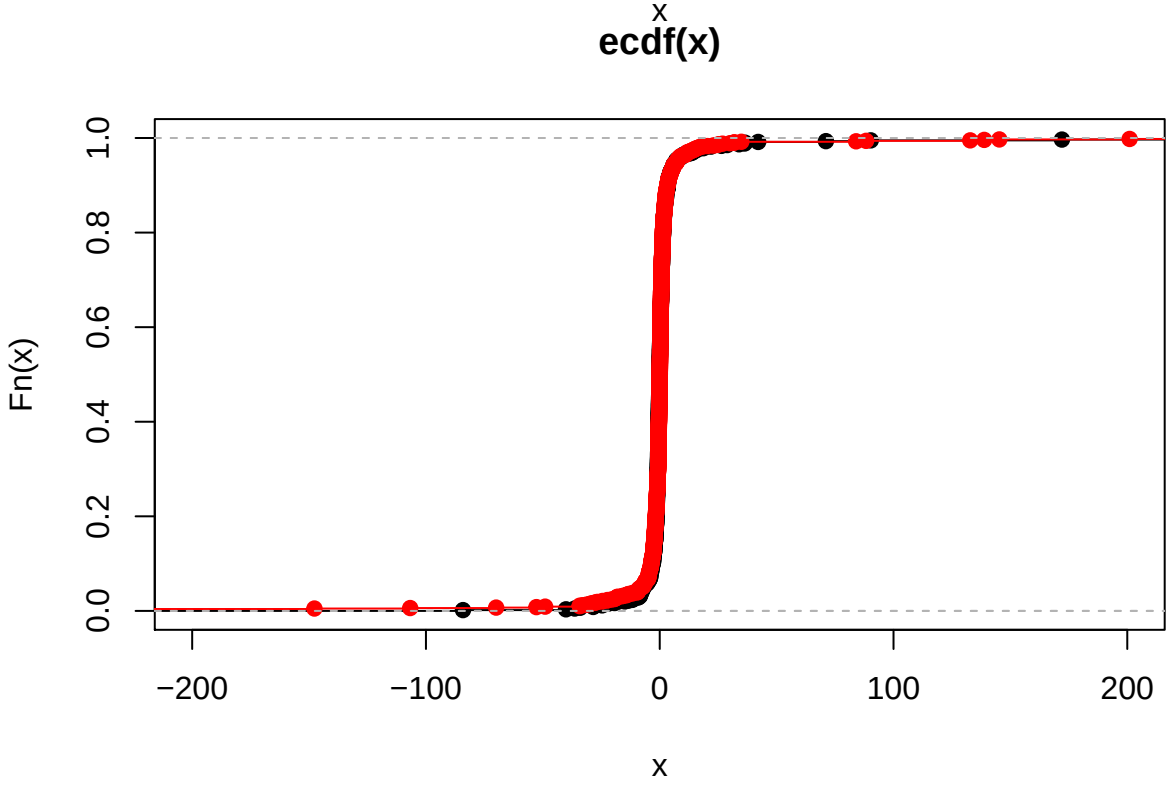
```
s <- seq(2, 1000, by=200)
for(n in s){
  x <- rcauchy(n)
```

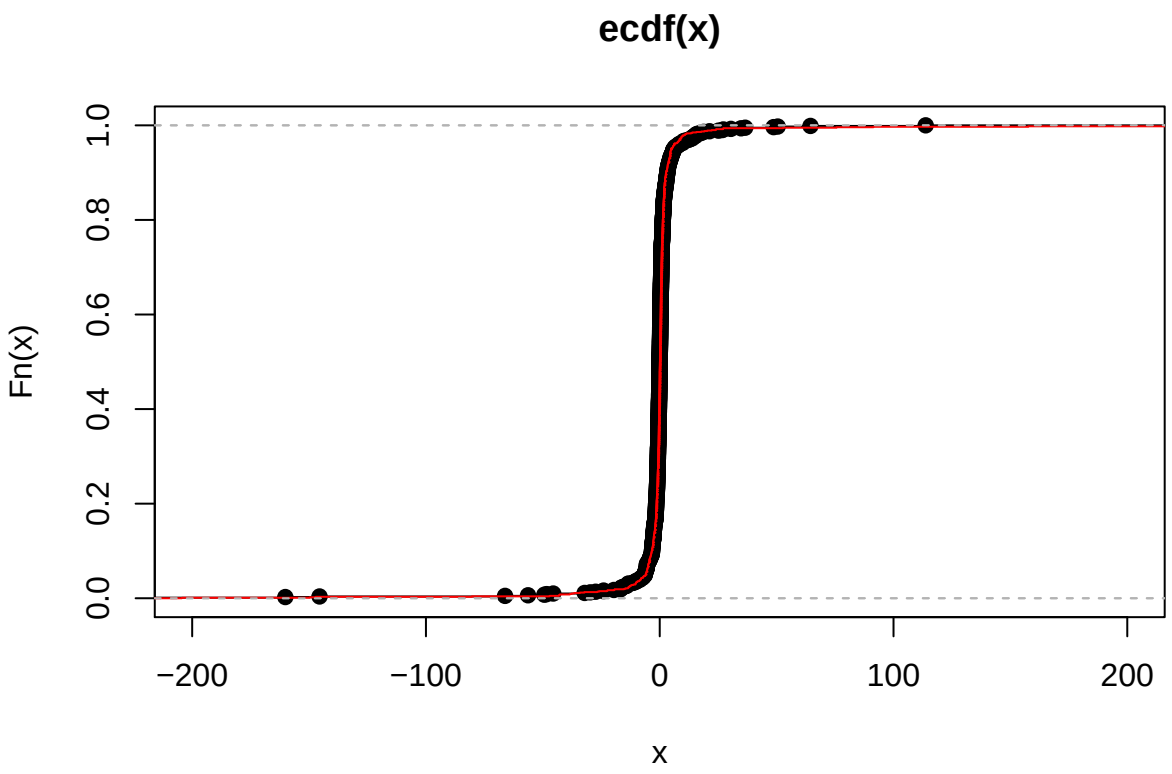
```
  y <- replicate(1000,mean(rcauchy(n)))
  plot.ecdf(x,xlim=c(-200,200))
  lines(ecdf(y),xlim=c(-200,200),col='red')
}
```






ecdf(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−200
−100
0
100
200
x


ecdf(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−200
−100
0
100
200
x


ecdf(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
−200
−100
0
100
200
x