How to calculate cumulative distribution in R?

Asked 7 years, 6 months ago Active 2 years, 7 months ago Viewed 139k times

23

votes

★

20

Locked. This question and its answers are locked because the question is off-topic but has historical significance. It is not currently accepting new answers or interactions.

I need to calculate the cumulative distribution function of a data sample.

Is there something similar to hist() in R that measure the cumulative density function?

I have tries ecdf() but i can't understand the logic.

r distributions cdf

edited Jun 22 '12 at 8:25
Jeromy Anglim
39.6k 22 132 234

asked Jun 21 '12 at 8:20
emanuele
1,686 2 15 32

comments disabled on deleted / locked posts / reviews

4 Answers

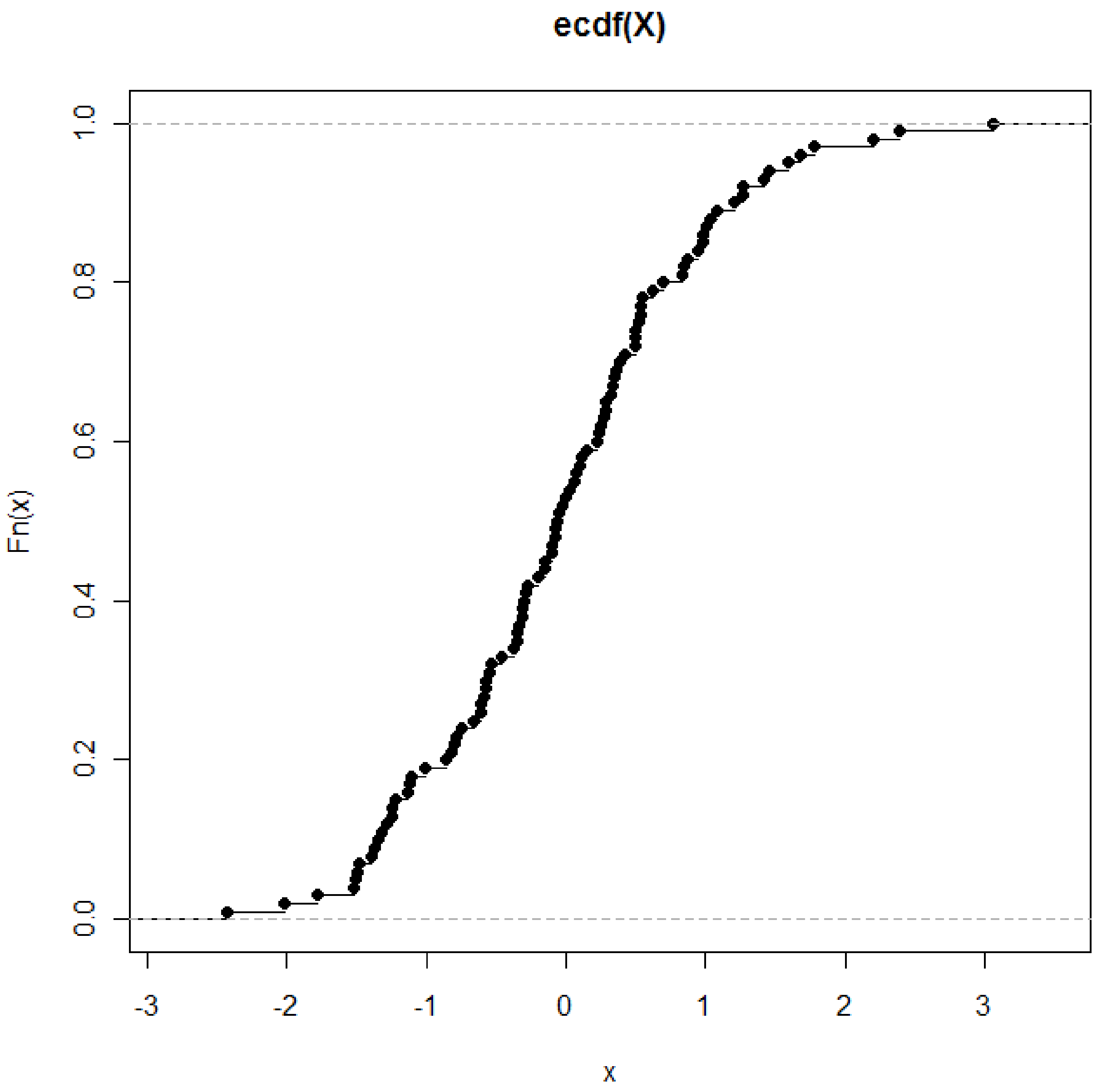
32

votes

✓

The `ecdf` function applied to a data sample returns a *function* representing the empirical cumulative distribution function. For example:

```
> X = rnorm(100) # X is a sample of 100 normally distributed random variables
> P = ecdf(X)    # P is a function giving the empirical CDF of X
> P(0.0)         # This returns the empirical CDF at zero (should be close to 0.5)
[1] 0.52
> plot(P)        # Draws a plot of the empirical CDF (see below)
```



If you want to have an object representing the empirical CDF evaluated at specific values (rather than as a function object) then you can do

```
> z = seq(-3, 3, by=0.01) # The values at which we want to evaluate the empirical CDF
> p = P(z)                # p now stores the empirical CDF evaluated at the values in z
```

Note that `p` contains at most the same amount of information as `P` (and possibly it contains less) which in turn contains the same amount of information as `X` .

edited Jun 21 '12 at 14:55

answered Jun 21 '12 at 8:48
Chris Taylor
3,104 1 22 26

Yes i know, but how is it possible to access the values of ecdf? this is a mystery for me. – emanuele Jun 21 '12 at 8:50

2 If you want its value at `x` you simply write `P(x)` . Note that `x` can be a vector (see the last couple of sentences of my answer.) – Chris Taylor Jun 21 '12 at 8:54

@ChrisTaylor The correct terminology is empirical cumulative distribution function not density function. – Michael R. Chernick Jun 21 '12 at 14:51

1
vote

What you appear to need is this to get the acumulated distribution (probability of get a value <= than x on a sample), ecdf returns you a function, but it appears to be made for plotting, and so, the argument of that function, if it were a stair, would be the index of the tread.

You can use this:

```
acumulated.distrib= function(sample,x){
    minors= 0
    for(n in sample){
        if(n<=x){
            minors= minors+1
        }
    }
    return (minors/length(sample))
}

mysample = rnorm(100)
acumulated.distrib(mysample,1.21) #1.21 or any other value you want.
```

Sadly the use of this function is not very fast. I don't know if R has a function that does this returning you a function, that would be more efficient.

answered Jun 1 '15 at 3:00
Casas
11 1

3 You seem to mix up the ECDF with its inverse. `R` does, indeed, compute the ECDF: its argument is a potential value of the random variable and it returns values in the interval $[0, 1]$. This is readily checked. For instance, `ecdf(c(-1,0,3,9))(8)` returns `0.75`. A generalized inverse of the ECDF is the quantile function, implemented by `quantile` in `R`. – whuber ♦ Jun 1 '15 at 16:19

1
vote

I always found `ecdf()` to be a little confusing. Plus I think it only works in the univariate case. Ended up rolling my own function for this instead.

First install data.table. Then install my package, mltools (or just copy the empirical_cdf() method into your R environment.)

Then it's as easy as

```
# Load packages
library(data.table)
library(mltools)

# Make some data
dt <- data.table(x=c(0.3, 1.3, 1.4, 3.6), y=c(1.2, 1.2, 3.8, 3.9))
dt
     x   y
1: 0.3 1.2
2: 1.3 1.2
3: 1.4 3.8
4: 3.6 3.9
```

CDF of a vector

```
empirical_cdf(dt$x, ubounds=seq(1, 4, by=1.0))
   UpperBound N.cum  CDF
1:          1     1 0.25
2:          2     3 0.75
3:          3     3 0.75
4:          4     4 1.00
```

CDF of column 'x' of dt

```
empirical_cdf(dt, ubounds=list(x=seq(1, 4, by=1.0)))
   x N.cum  CDF
1: 1     1 0.25
2: 2     3 0.75
3: 3     3 0.75
4: 4     4 1.00
```

CDF of columns 'x' and 'y' of dt

```
empirical_cdf(dt, ubounds=list(x=seq(1, 4, by=1.0), y=seq(1, 4, by=1.0)))
    x y N.cum  CDF
 1: 1 1     0 0.00
 2: 1 2     1 0.25
 3: 1 3     1 0.25
 4: 1 4     1 0.25
 5: 2 1     0 0.00
 6: 2 2     2 0.50
 7: 2 3     2 0.50
 8: 2 4     3 0.75
 9: 3 1     0 0.00
10: 3 2     2 0.50
11: 3 3     2 0.50
12: 3 4     3 0.75
13: 4 1     0 0.00
14: 4 2     2 0.50
15: 4 3     2 0.50
16: 4 4     4 1.00
```

edited Nov 28 '16 at 2:50

answered Nov 21 '16 at 19:14
Ben
1,173 1 11 26

1
vote

friend, you can read the code on this blog.

```
sample.data = read.table ('data.txt', header = TRUE, sep = "\t")
cdf <- ggplot (data=sample.data, aes(x=Delay, group =Type, color = Type)) + stat_ecdf()
cdf
```

more details can be found on following link:

r cdf and histogram

edited May 19 '17 at 11:14
Rudy Yuan
3 2

answered Mar 31 '16 at 2:27
CrossWorld2
11 2