

Empirical Distributions and CDFs:

Along with Quantiles Based on the Empirical CDF

Let's generate 5 numbers from a normal distribution with a mean of 0 and standard deviation of 2. Since these values were randomly generated the values you get will be different from the following:

```
> n <- 5
> y <- rnorm(n, mean=0, sd=2)
> round(sort(y),3)
[1] -2.122 -0.742 -0.259  0.139  0.311
```

In the code the values were sorted. let's call these $y_{(1)} = -2.122, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)} = 0.311$.

- What is $\hat{\Pr}(0, 1]$?

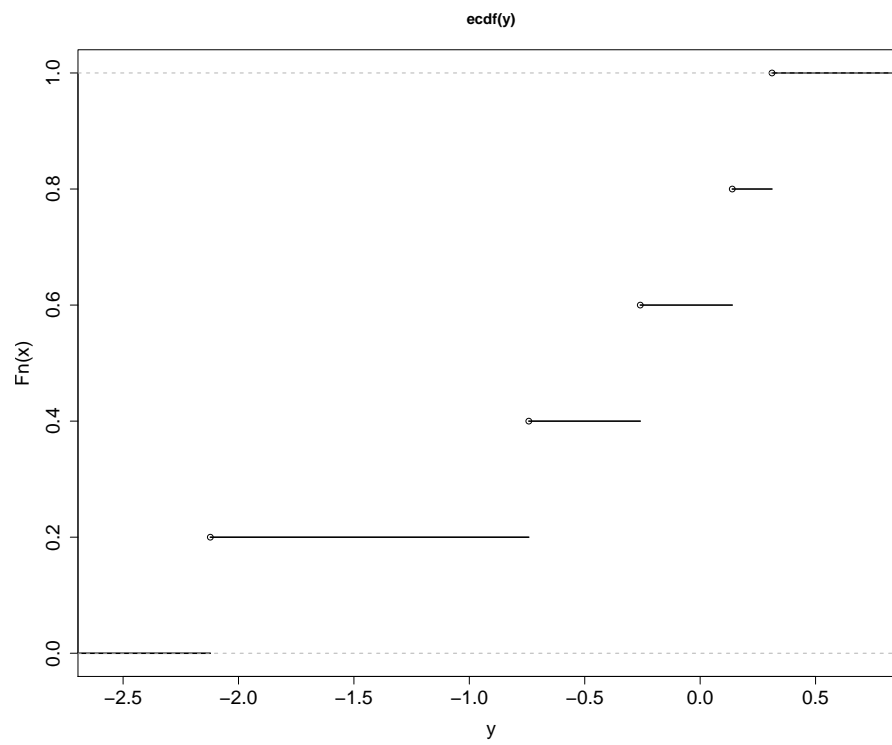
$$\begin{aligned}\hat{\Pr}(0, 1] &= \#(0 < y_i \leq 1)/n \\ &= 2/5 = 0.4\end{aligned}$$

```
> a <- 0
> b <- 1
> length(y[y>a & y<=b])/n
[1] 0.4
```

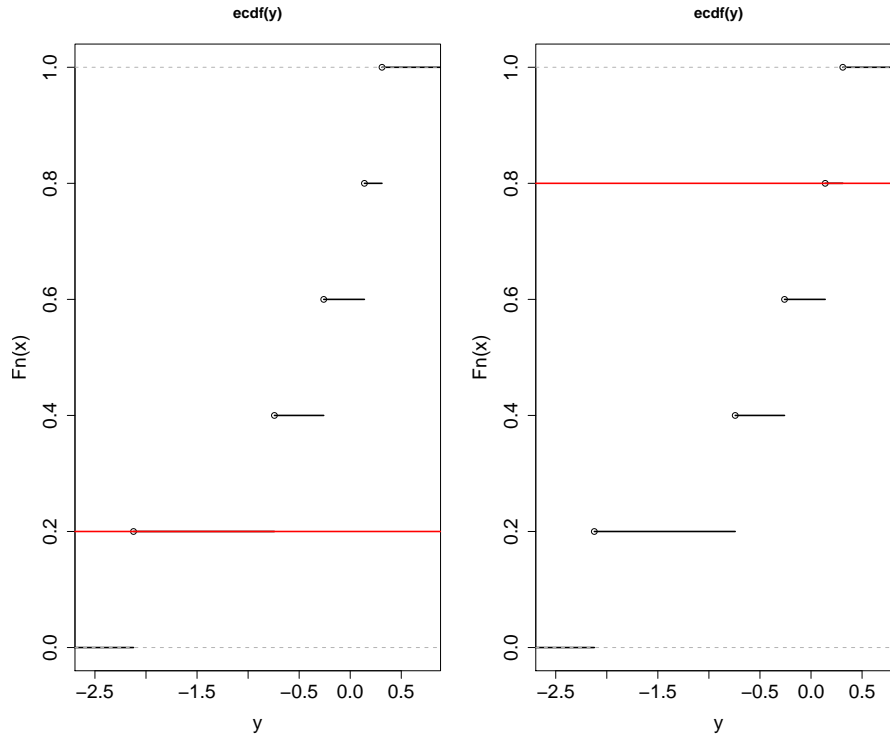
The empirical CDF is defined as $\hat{F}(y) = \#(y_i \leq y)/n = \hat{\Pr}(-\infty, y]$. Thus we have:

$$\begin{aligned}\hat{\Pr}(-\infty, y_{(1)} = -2.122] &= 1/5 = 0.20 \\ \hat{\Pr}(-\infty, y_{(2)} = -0.742] &= 2/5 = 0.40 \\ \hat{\Pr}(-\infty, y_{(3)} = -0.259] &= 3/5 = 0.60 \\ \hat{\Pr}(-\infty, y_{(4)} = 0.139] &= 4/5 = 0.80 \\ \hat{\Pr}(-\infty, y_{(5)} = 0.311] &= 5/5 = 1.00\end{aligned}$$

```
> plot(ecdf(y), lwd=2, cex.axis=1.25, cex.lab=1.25)
```



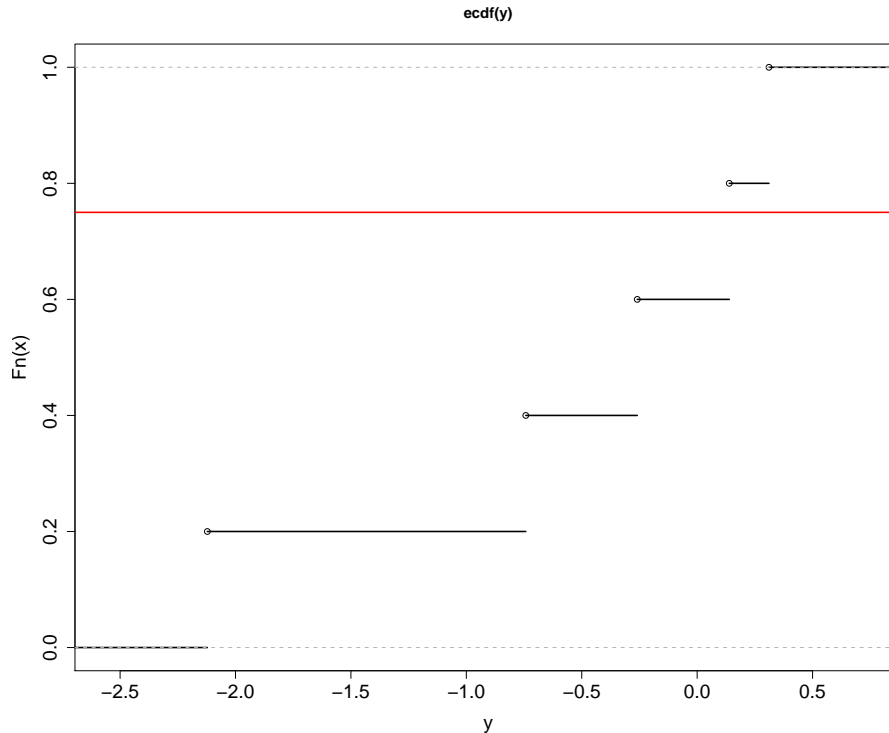
- What is the sample (empirical) 20% quantile (i.e. $\hat{q}_{(0.2)}$)?
- What is the sample (empirical) 80% quantile (i.e. $\hat{q}_{(0.8)}$)?



One approach to determine the quantiles is through the empirical distribution function above. Thus we want to use the associated empirical CDF of the order statistics $y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}$ as the quantiles. In R we have to set 'type=4'. Try other types, including the default to see the difference. So we will calculate the quantiles through examining $n \times p$:

- $5 \times 0.2 = 1$, so take $y_{(1)} = -2.122$.
- $5 \times 0.8 = 4$, so take $y_{(4)} = 0.139$.

```
> quantile(y, prob=0.20, type=4)
 20%
-2.122108
> quantile(y, prob=0.80, type=4)
 80%
0.1391126
```



- What is the sample (empirical) 75% quantile (i.e. $\hat{q}_{(0.75)}$)?
 - Now $n \times p = 5 \times 0.75 = 3.75$. So One approach is to use a linear combination between $y_{(3)}$ and $y_{(4)}$:

$$\hat{q}_{(0.75)} = (1 - \gamma) \times y_{(3)} + \gamma \times y_{(4)},$$

where $\gamma = \frac{0.80 - 0.75}{0.80 - 0.60} = 0.75$, or the decimal portion of $n \times p = 5 \times 0.75 = 3.75$.

```
> # the length between F(y(3)) and F(y(4))
> l.y4.y3 <- 0.8 - 0.6
>
> # the length between F(y[3]) and 0.75
> l.yq.y3 <- 0.75 - 0.6
>
> gamma <- l.yq.y4/l.y5.y4

> # thus the 75% quantile is gamma*y(0.60) + (1-gamma)*y(0.80)
> (1-gamma)*y.sort[3] + gamma*y.sort[4]
[1] 0.0395
>
> # or
> quantile(y, prob=0.75, type=4)
 75%
0.0395
```

- This all seems reasonable, but now what is the median (i.e. $\hat{q}_{(0.5)}$)?

– Now $n \times p = 5 \times 0.5 = 2.5$, so $\hat{q}_{(0.5)} = (1 - 0.5) \times y_{(2)} + 0.5 \times y_{(3)} = 0.5(-0.742) + 0.5(-0.259) = -0.5005$.

```
> quantile(y, prob=0.5, type=4)
 50%
-0.5005
```

What do you think of this as an answer? Doesn't $y_{(3)}$ appear to be the median as was stated in the notes on page 9? For example the default function in R gives the answer you would expect:

```
> quantile(y, prob=0.5)
 50%
-0.259
```

So there are different ways to consider the problem lying in the discreteness of the observations. Look at `help(quantile)` in R. Also look at the following journal article ‘Sample Quantiles in Statistical Packages’ by Hyndman and Fan; there is a link on the website. **For the course use either the default function in R or ‘type=4’.**