Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#224 Quantiles

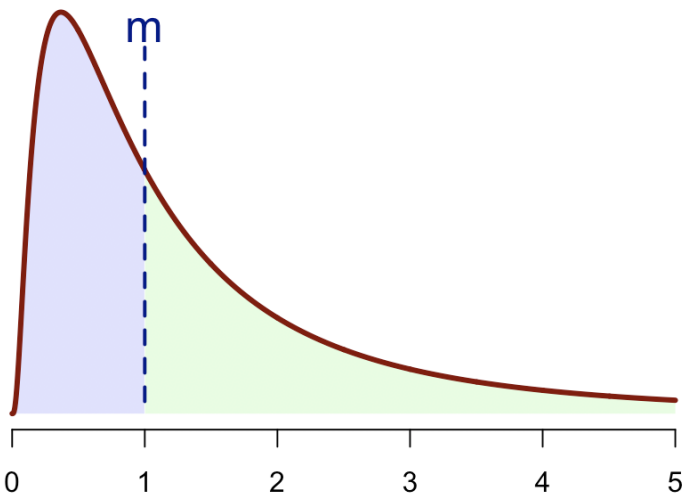
été 2020

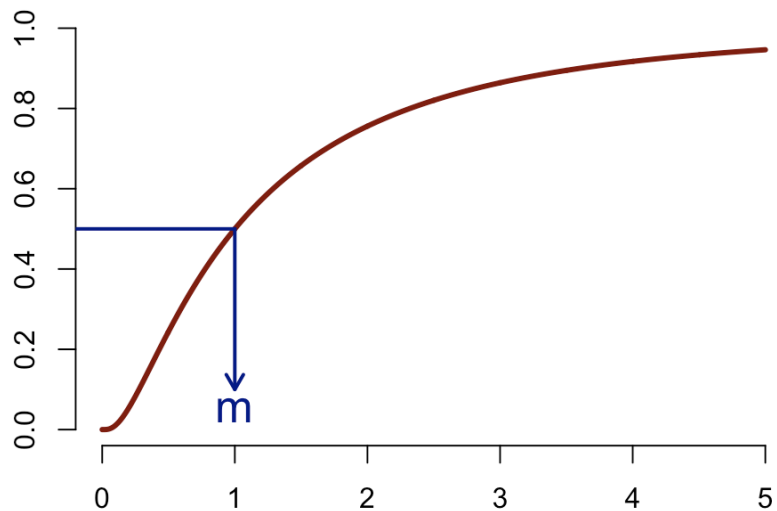
Median

The median is a value m such that

$$\mathbb{P}(X \leq m) \geq \frac{1}{2} \text{ and } \mathbb{P}(X \geq m) \geq \frac{1}{2}$$

In a nutshell, *the median is the value separating the higher half from the lower half of a data sample* (see https://en.wikipedia.org)





Median

Let $\boldsymbol{y} \in \mathbb{R}^n$, $\text{median}[\boldsymbol{y}] \in \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \frac{1}{n} |\underbrace{y_i - m}_{\varepsilon_i}| \right\}$

It is the empirical version of

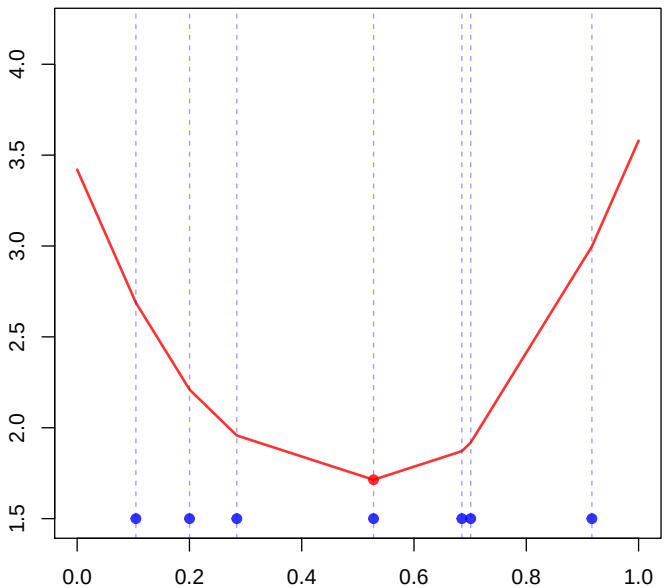
$$\text{median}[Y] \in \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \int |\underbrace{y - m}_{\varepsilon}| dF(y) \right\} = \underset{m \in \mathbb{R}}{\text{argmin}} \left\{ \mathbb{E}\left[\| \underbrace{Y - m}_{\varepsilon} \|_{\ell_1} \right] \right\}$$

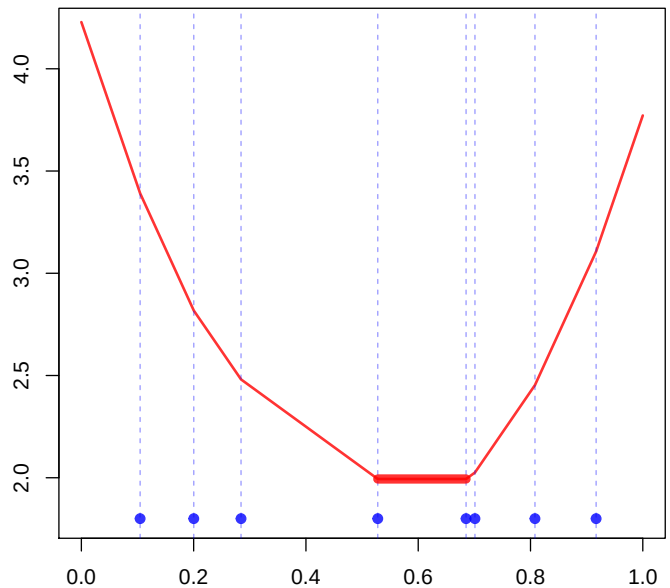
where Y is a random variable, $\mathbb{P}[Y \leq \text{median}[Y]] \geq \frac{1}{2}$ and $\mathbb{P}[Y \geq \text{median}[Y]] \geq \frac{1}{2}$.

See Boscovich (1757) De Litteraria expeditione per pontificiam ditionem ad dimetiendos duos meridiani and Laplace (1793) Sur quelques points du système du monde.

Median and Minimization

Sketch of proof: Let $h(x) = \sum_{i=1}^{n} |x - y_i|$





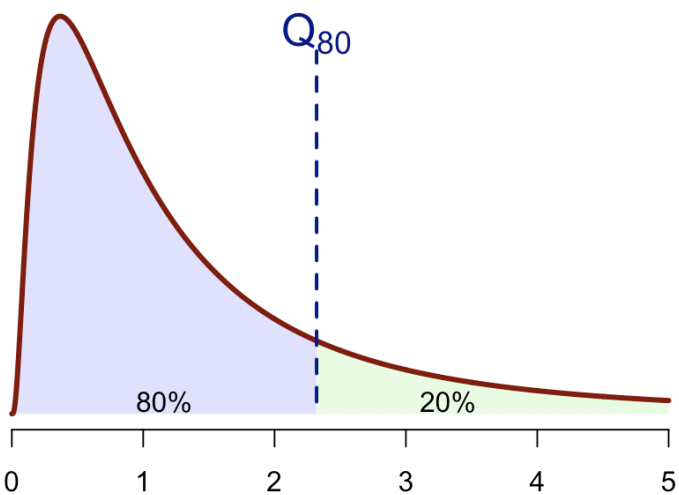
The median is unique when n is odd, not when n is even...

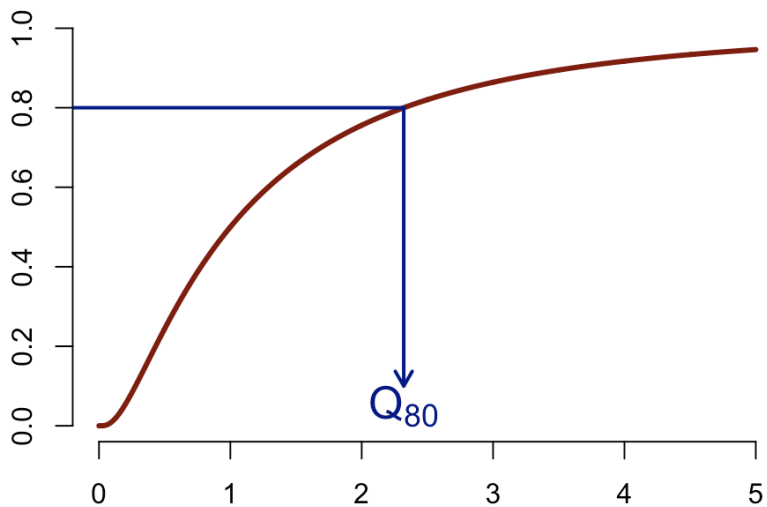
Quantile

The quantile of level $\alpha \in (0, 1)$ is a value Q_α such that

$$\mathbb{P}(X \leq Q_\alpha) \geq \alpha \text{ and } \mathbb{P}(X \geq Q_\alpha) \geq 1 - \alpha$$

Hence *the median is the value separating the higher* $1 - \alpha$*% from the lower* α*% of a data sample*





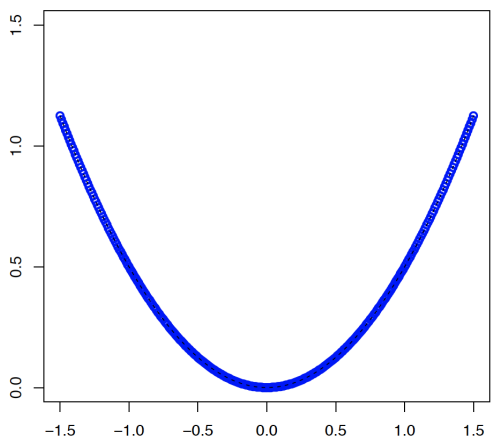
Quantile

$$Q_\alpha \in \text{argmin}\left\{ \sum_{i=1}^{n} \omega_\alpha(\varepsilon_i) \big| \underbrace{y_i - q_i}_{\varepsilon_i} \big| \right\} \text{ where } \omega_\alpha(\epsilon) = \begin{cases} 1-\alpha \text{ if } \epsilon \leq 0 \\ \alpha \text{ if } \epsilon > 0 \end{cases}$$

When $\alpha = 1/2$ we have the median,

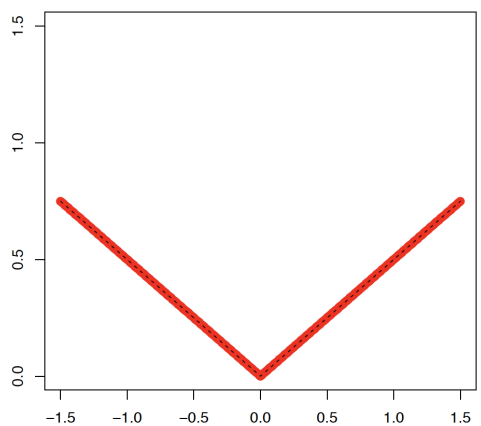
average

$$\sum_{i=1}^{n} \frac{1}{2}(y_i - m)^2$$



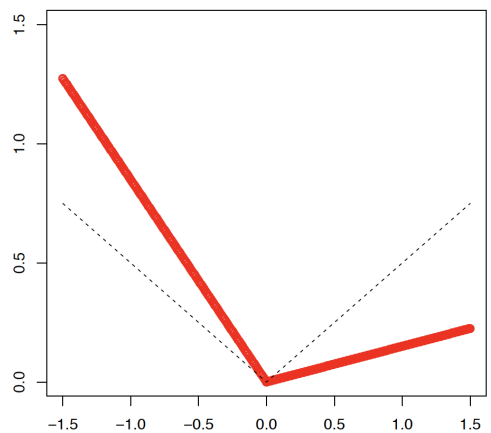
median

$$\sum_{i=1}^{n} \frac{1}{2}|y_i - m|$$



quantile

$$\sum_{i=1}^{n} \omega_i |y_i - m|$$



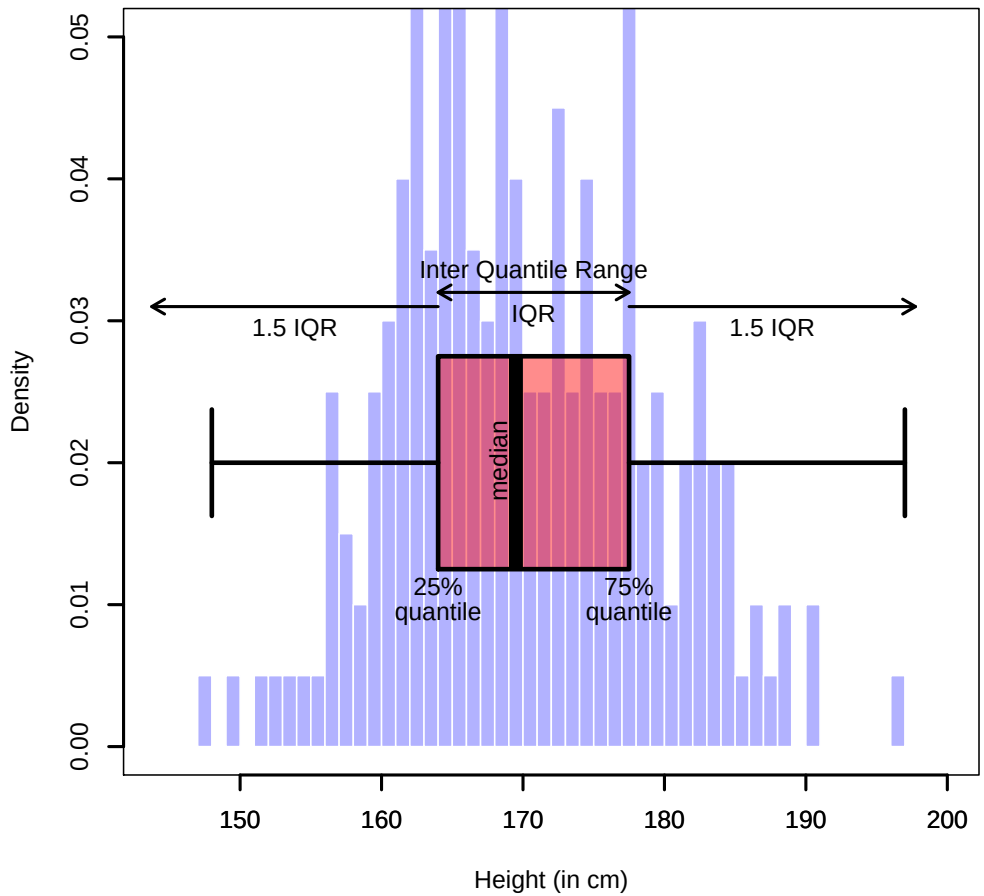
Quantiles and Box-Plot

Box plot, see Tukey's Exploratory Data Analysis

The box corresponds to the $(25\% - 75\%)$ quantile

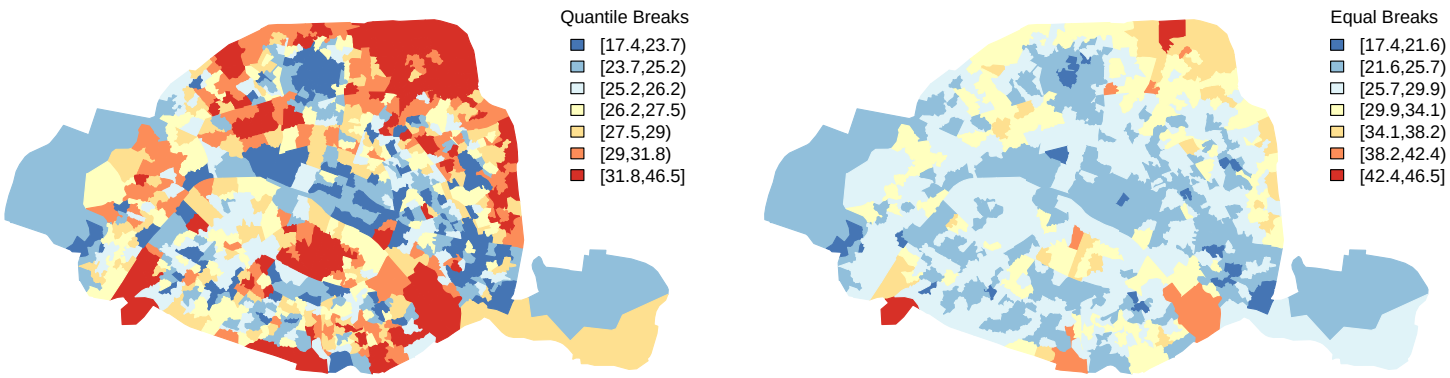
The ends of the whiskers are quantiles $\pm 1.5 IQR$,

$IQR = Q_{75\%} - Q_{25\%}$



Quantiles and color Palette

It is possible to use colors to visualize the scale of y (with a appropriate gradient)
y : proportion of population below 24 years old



see Error on Choroplethic Maps: Definition, Measurement, Reduction