$x = (x_1, \dots, x_n)$, $x_i \in \mathbb{R}$

media, varianza, scarto quadratico medio

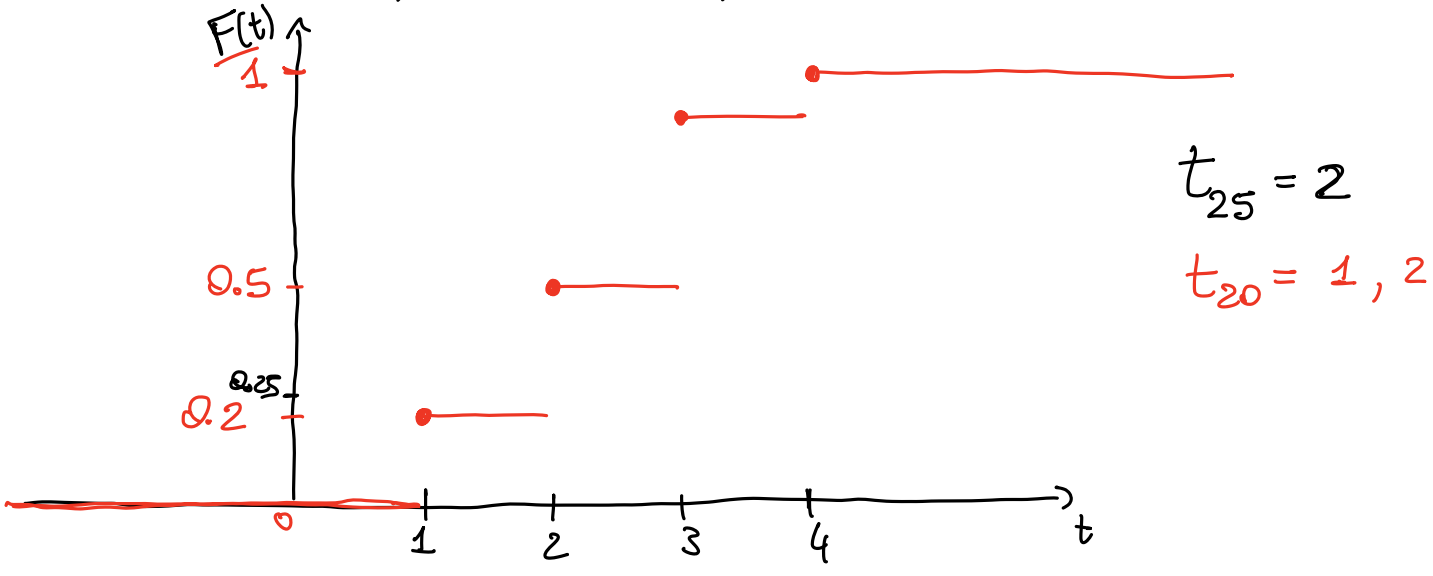
Def Funzione di ripartizione $F: \mathbb{R} \to \mathbb{R}$

$$F(t) := \frac{\#\{x_i : x_i \le t\}}{n} \in [0,1]$$

Proprietà

- $\lim_{t \to -\infty} F(t) = 0$, $\lim_{t \to +\infty} F(t) = 1$
- F è debolmente crescente
  ($se\ t_1 < t_2$ allora $F(t_1) \le F(t_2)$)
- F è continua da destra
  ($\lim_{t \to t_0^+} F(t) = F(t_0)$)

Es $x = (1, 1, 2, 2, 2, 3, 3, 3, 3, 4)$



Per $x = (x_1, \dots, x_n)$, $F(t)$ è costante a tratti con salti in $t \in \{x_1, \dots x_n\}$ di grandezza

uguale a $\frac{1}{n} \#\{x_i : x_i = \bar{t}\} \quad \forall\, \bar{t} \in \{x_1, \dots, x_n\}$

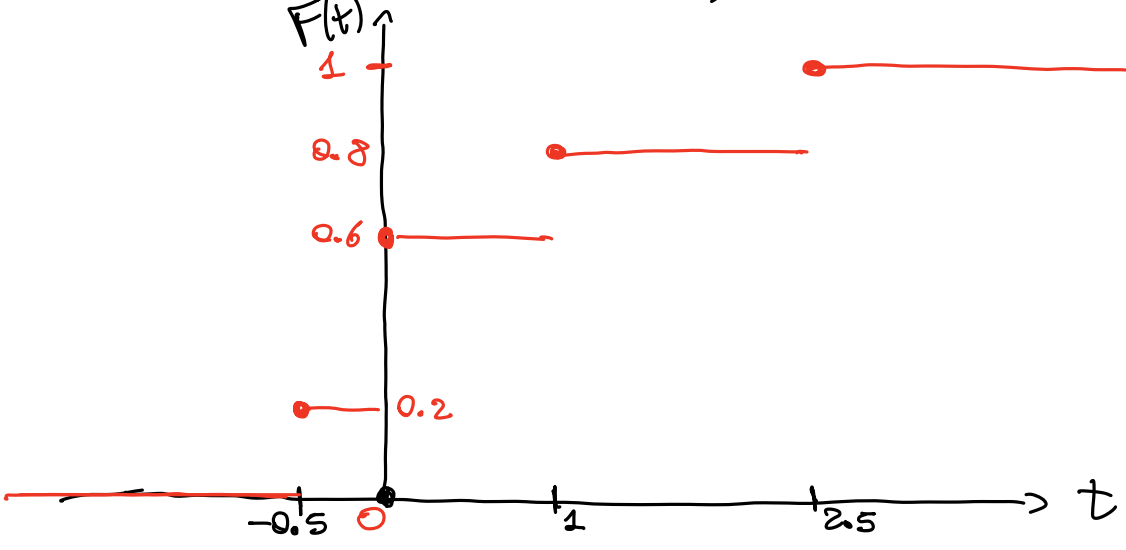
Def Dato $0 < k < 100$ si chiama
k-simo percentile un valore $\boxed{t_k}$ tale che:

- almeno il $k\%$ dei dati è minore o uguale a t_k;
- almeno il $(100-k)\%$ dei dati è maggiore o uguale a t_k.

OSS Se $\beta = \frac{k}{100} \in (0,1)$ il k-simo percentile si chiama anche **β-quantile**

Se $\beta = 0.25$, primo quantile
$\beta = 0.5$, mediana (secondo quantile)
$\beta = 0.75$, terzo quantile

ES $x = (0, 1, 2.5, -0.5, 0)$



$\beta = 0.25 \quad t_\beta = 0$
$\beta = 0.5 \quad t_\beta = 0$
$\beta = 0.75 \quad t_\beta = 1$

Dati multipli $x = (x_1, \dots, x_n)$

$y = (y_1, \dots, y_n)$

Def Si chiama COVARIANZA TRA x e y

cov. CAMPIONARIA $\quad cov(x,y) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

cov. EMPIRICA $\quad cov_e(x,y) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

OSS • $cov(x,x) = var(x)$

• $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}$

Def Se $\sigma(x) \neq 0$, $\sigma(y) \neq 0$, si definisce COEFFICIENTE DI CORRELAZIONE

$$r(x,y) = \frac{cov(x,y)}{\sigma(x)\,\sigma(y)} = \frac{cov_e(x,y)}{\sigma_e(x)\,\sigma_e(y)}$$

$$\left(= \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \right)$$

Prop $r(x,y) \in [-1, 1]$

dim Applicare disuguaglianza di Schwartz

$$\left[|\textstyle\sum a_i b_i| \leq \left(\sum a_i^2\right)^{1/2} \cdot \left(\sum b_i^2\right)^{1/2} \right]$$

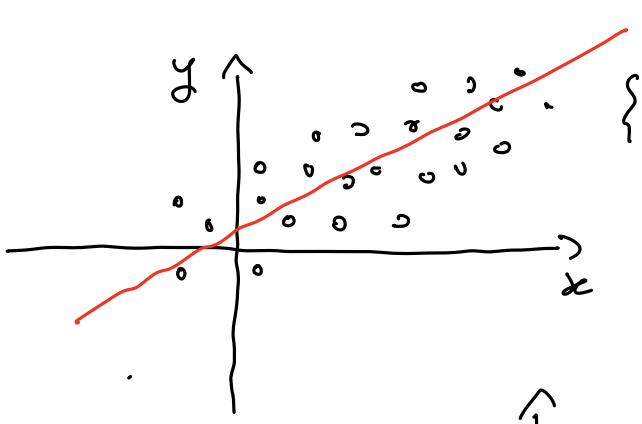
$$\left| \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right| \leq \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{1/2} \qquad \square$$

- Se $|r(x,y)| \gtrsim 0.7$ allora x e y sono CORRELATI

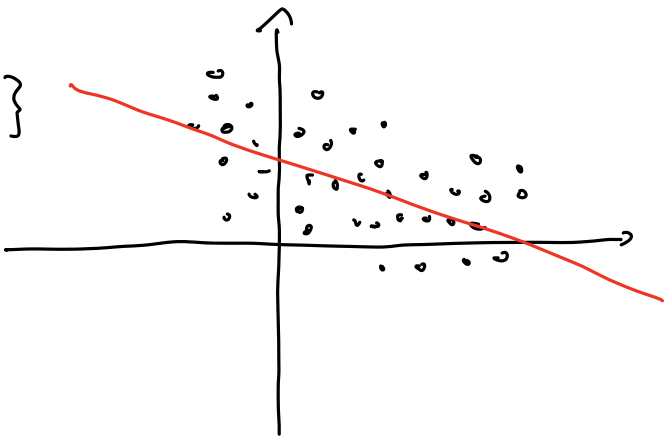
  $|r(x,y)| \leq 0.3$ allora x e y sono SCORRELATI

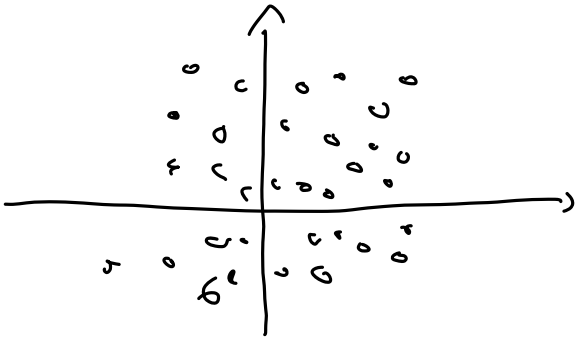
- se $y_i = b x_i + a \quad \forall i$, con $a, b \in \mathbb{R}$

  allora $r(x,y) = \text{segno}(b) = \pm 1$



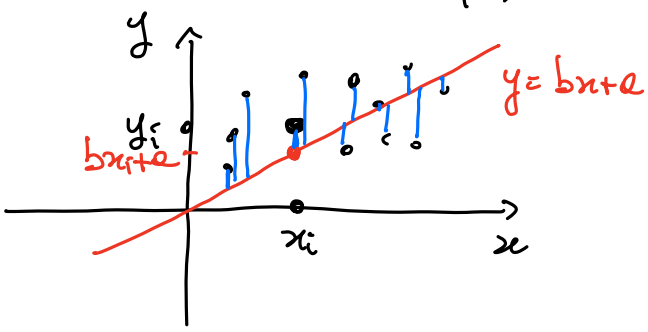
$\{(x_i, y_i)\}$





TEOREMA Dati $x = (x_1, \dots x_n)$ e $y = (y_1, \dots y_n)$ con $\sigma(x) \neq 0$ e $\sigma(y) \neq 0$, definiamo la funzione

$$Q(a,b) := \sum_{i=1}^{n} (y_i - b x_i - a)^2$$



La funzione $Q(a,b)$ ha minimo nel punto (a^*, b^*) con

$$b^* = \frac{\text{cov}(x,y)}{\text{var}(x)}, \quad a^* = \bar{y} - b^* \bar{x}$$

e la retta $y = b^* x + a^*$ si chiama RETTA DI REGRESSIONE. Inoltre

$$Q(a^*, b^*) = \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right] (1 - r(x,y)^2).$$