Statistica e probabilità

- Statistica descrittiva

Siano $x_1, \dots, x_N$ $(= x)$ un numero finito di dati.

Frequenza assoluta di un dato $x_i := \# \{ k \,/\, x_k = x_i \}$

$\uparrow$ cardinalità

<u>Esempio</u> $x = \{$ Luglio, Marzo, Febbraio, Aprile, Gennaio, Agosto, Marzo, Aprile, Dicembre, Novembre, Giugno $\}$

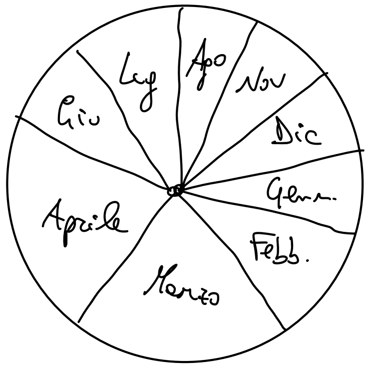
Frequenza assoluta di
Gennaio = 1 , Settembre = 0
Febbraio = 1 , Ottobre = 0
Marzo = 2 , Novembre = 1
Aprile = 2 , Dicembre = 1
Maggio = 0
Giugno = 1
Luglio = 1
Agosto = 1

- Frequenza relativa di un dato x_i $= \dfrac{\text{Frequenza assoluta di } x_i}{N}$

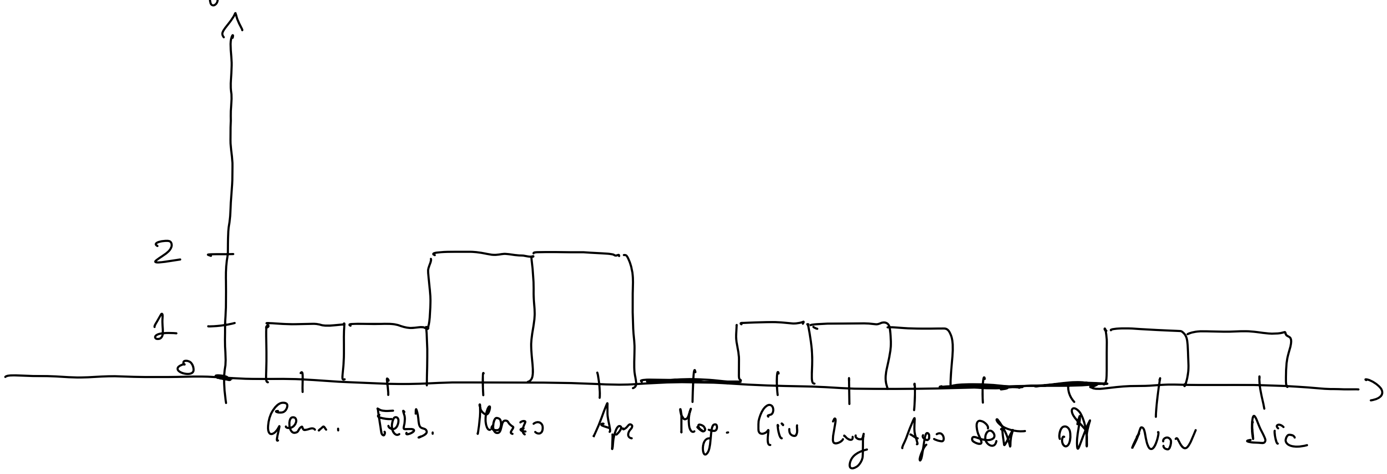
<u>ES</u> Frequenza relativa di Gennaio $= \frac{1}{11} \sim 9\%$
Febbraio $= \frac{1}{11} \sim 9\%$
Marzo $= \frac{2}{11} \sim 18\%$
$\vdots$

<u>Rappresentazioni</u>

Diagramma a Torta

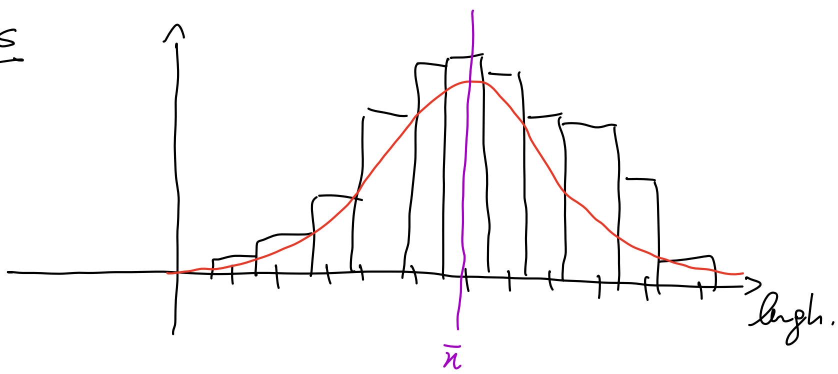


Istogramma



Supponiamo di avere dati quantitativi (es altezze, giorno di nascita, ...)

ES



Normale

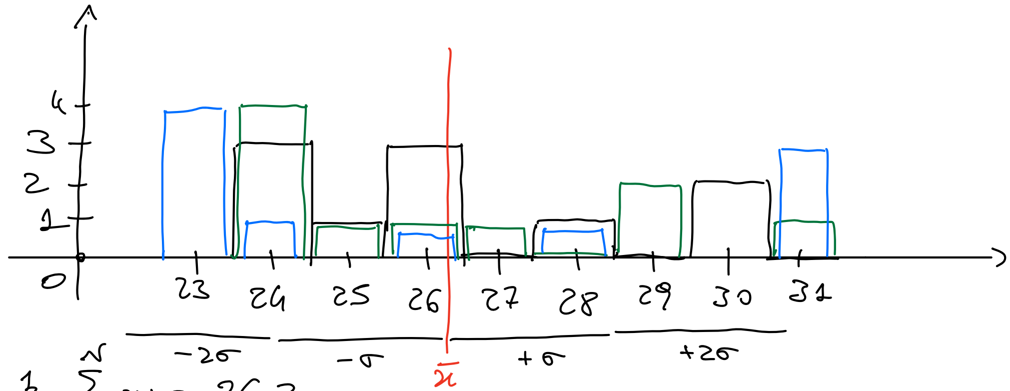
"curva a campana"

Def Dato un campione $x = \{x_1, \dots, x_N\}$ di N dati, si chiama MEDIA di x, il numero

$$\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} (x_1 + x_2 + \dots + x_N)$$

OSS $\min_{a \in \mathbb{R}} \sum_{i=1}^{N} (x_i - a)^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2$

ES $\quad x = \{25, 24, 28, 30, 26, 24, 26, 30, 24, 26\}$ $\qquad N = 10$



$$\bar{x} = \frac{1}{10} \sum_{i=1}^{N} x_i = 26,3$$

$x = \{24, 24, 29, 31, 25, 24, 27, 29, 24, 26\}$, $\bar{x} = 26,3$

$x = \{23, 23, 31, 31, 24, 23, 28, 31, 23, 26\}$, $\bar{x} = 26,3$

Def Dato un campione $x = \{x_1, \dots, x_N\}$ con media $\bar{x}$, si chiama VARIANZA (empirica) DI x il numero

$$var_e(x) := \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Si chiama SCARTO QUADRATICO MEDIO DI x (o DEVIAZIONE STANDARD DI x) la quantità

$$\sigma(x) := \sqrt{var(x)}$$

ES

$x = \{25, 24, 28, 30, 26, 24, 26, 30, 24, 26\}$

$x = \{24, 24, 29, 31, 25, 24, 27, 29, 24, 26\}$

$x = \{23, 23, 31, 31, 24, 23, 28, 31, 23, 26\}$

$\bar{x} = 26,3$

$$var(x) = \begin{cases} 4,81 \\ 6,01 \\ 11,81 \end{cases} \qquad \sigma(x) = \begin{cases} 2,19 \\ 2,45 \\ 3,44 \end{cases}$$

OSS $\quad var(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) =$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \underbrace{\frac{2}{N}\sum_{i=1}^{N} x_i\bar{x}}_{\substack{= \\ \frac{2\bar{x}}{N}\sum_{i=1}^{N} x_i \\ = \\ 2\bar{x}^2}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\bar{x}^2}_{\substack{= \\ \frac{1}{N}\cdot N\bar{x}^2 \\ = \\ \bar{x}^2}} =$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

Prop (Disuguaglianza di Chebyshev) Dato $x=\{x_1,\dots,x_N\}$ con media $\bar{x}$ e varianza $var(x)$, vale che per ogni $d>0$

$$\frac{\#\{i \,/\, |x_i-\bar{x}|>d\}}{N} \le \frac{var(x)}{d^2}$$

$$\left(\text{equivalentemente} \quad \frac{\#\{i \,/\, |x_i-\bar{x}|\le d\}}{N} \ge 1-\frac{var(x)}{d^2}\right)$$

OSS $\quad d = \alpha\cdot\sigma(x)$, $\alpha>0$, $\quad \sigma(x)=\sqrt{var(x)}$

$$\frac{\#\{i \,/\, |x_i-\bar{x}|\le \alpha\,\sigma(x)\}}{N} \ge 1 - \frac{var(x)}{\alpha^2(\sigma(x))^2} = 1-\frac{1}{\alpha^2}$$

ES $\alpha=3$, Almeno 89% dei dati è a distanza $\le 3\sigma$ da $\bar{x}$.

OSS Per la distribuzione normale vale che:

almeno il 68% dei dati è a distanza $\le\sigma$ da $\bar{x}$

" " 95% " " " " " $\le 2\sigma$ "

" " 99,7 % " " " " " $\leq 3\sigma$ "

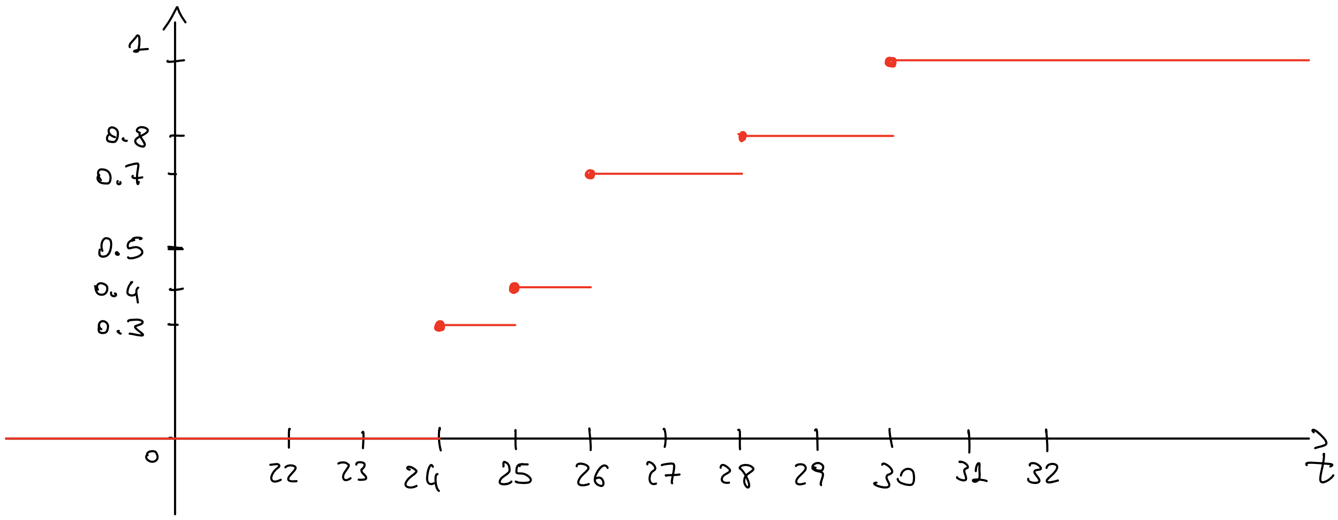
- Funzione di ripartizione $F(t)$, $t \in \mathbb{R}$, $F : \mathbb{R} \to \mathbb{R}$

$\mathrm{Im}\, F = [0,1]$.

Dato $x = \{x_1, \dots, x_N\}$,

$$F(t) := \frac{\#\{i \,/\, x_i \leq t\}}{N}$$

ES $x = \{25, 24, 28, 30, 26, 24, 26, 30, 24, 26\}$



$F(24) = \frac{1}{10} \#\{i \,/\, x_i \leq 24\} = \frac{3}{10}$

Def Dato $k \in (0,100)$, si chiama K-SIMO PERCENTILE DI $x = \{x_1, \dots, x_N\}$ il valore $t_k \in \mathbb{R}$ per cui:

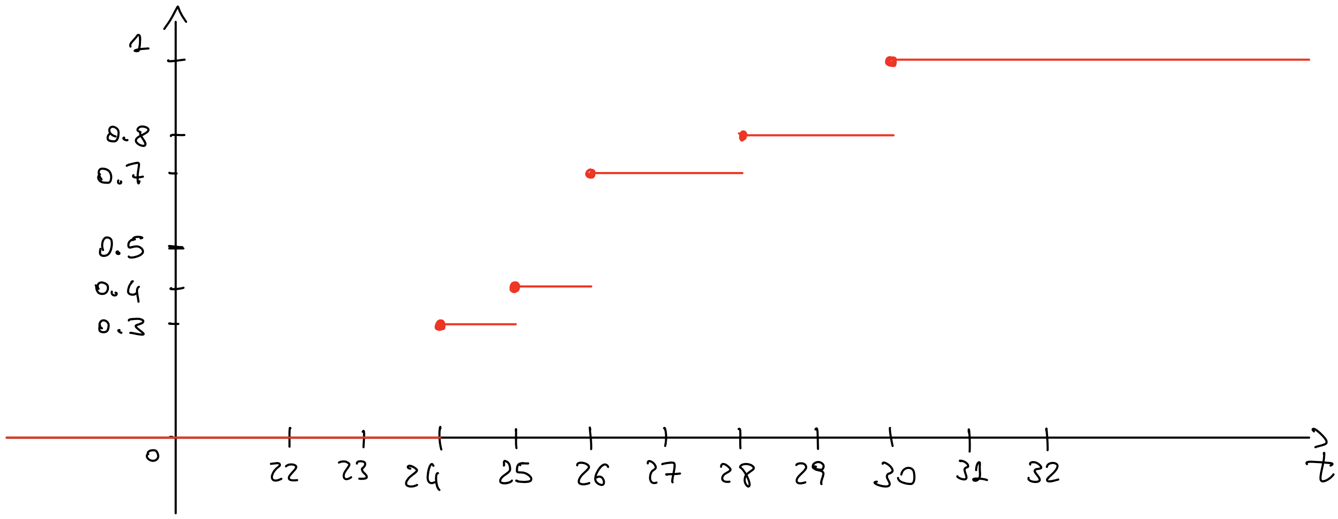
- almeno $\frac{k}{100} \cdot N$ dati sono $\leq t_k$
- almeno $\left(1 - \frac{k}{100}\right)N$ dati sono $\geq t_k$.

Se $k = 25$, il numero t_{25} si chiama primo quartile

" $k = 50$, " " t_{50} " mediana

" $k = 75$, " " t_{75} " terzo quartile

ES



La mediana è 26 (infatti ho 7 dati su 10 che sono ≤ 26 e ho 6 dati su 10 che sono ≥ 26)

Il 40-esimo percentile è 25,5 (infatti ho 4 dati su 10 che sono $\leq 25{,}5$ e ho 6 dati su 10 che sono $\geq 25{,}5$)

Il primo quartile è 24, il terzo quartile è 28.