CLAUDIO BONANNO

DIDATTICA / STATISTICA

STATISTICA

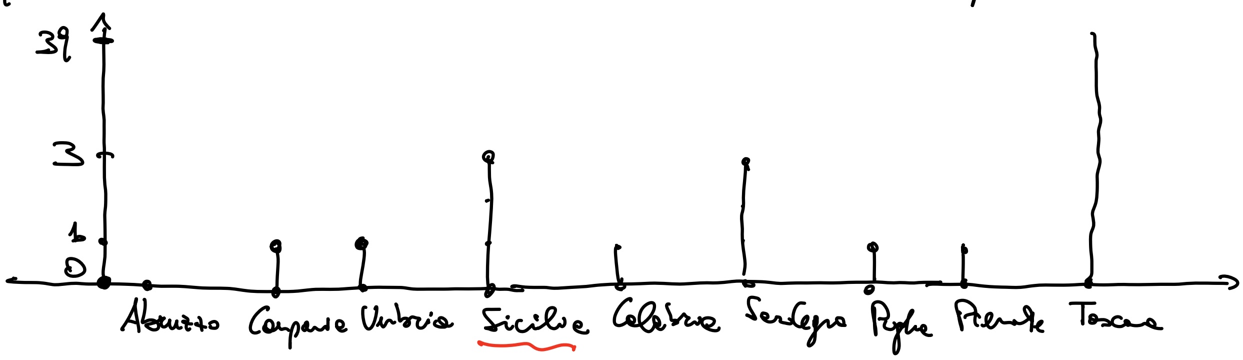
- Note del corso
- Ross, "Probabilità e statistica per l'ingegneria e le scienze"

Statistica descrittiva - Inferenza statistica

Statistica descrittiva

Dati → Qualitativi o Quantitativi

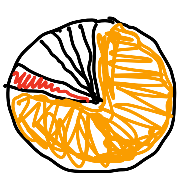
Frequenza (assoluta) = # delle volte con cui compare un dato



$$\text{Frequenza relativa} = \frac{\text{Freq. assoluta}}{\text{numero dei dati}} = \text{percentuale di comparsa del dato}$$

numero totale = 50

$$\text{Freq. relativa (Toscana)} = \frac{39}{50} = 0{,}78$$, percentuale 78 %



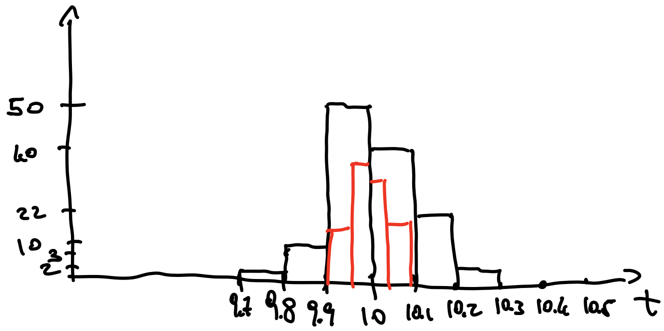
Sicilia $\simeq \frac{3}{50} = 0.06$

$\updownarrow$

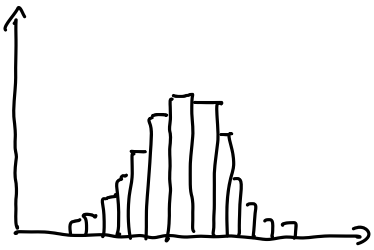
6 %

Quantitativi

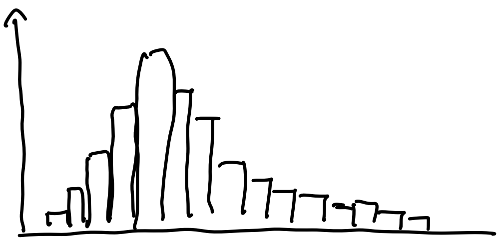
Diagramma a bastoni



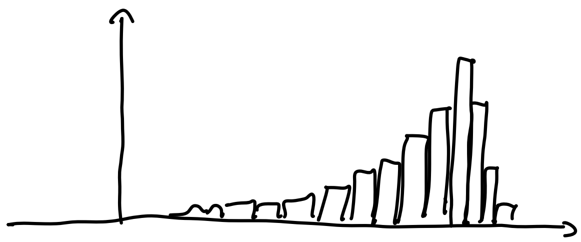
- Normale



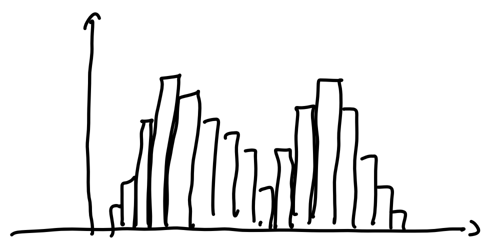
- Right-skewed



- Left-skewed



- Bimodale



Dati numerici

$x = (x_1, x_2, \dots, x_n)$ dati, $x_i \in \mathbb{R}$

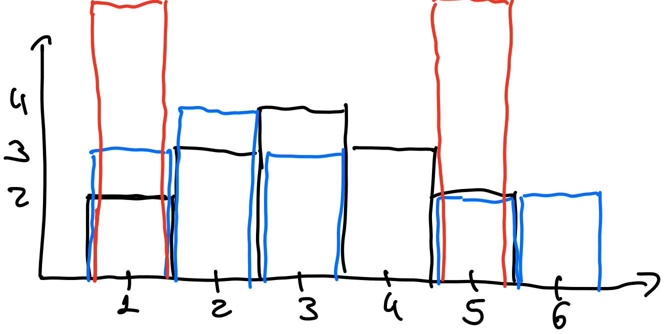
Def Media (empirica o campionaria)

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Esempio $x = (1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5)$, $n = 14$

$$\bar{x} = \frac{1}{14}(1 + 1 + \dots + 5 + 5) = \frac{42}{14} = 3, \quad var_e(x) = \frac{22}{14} \sim 1.571$$

$$\sigma_e(x) \sim 1.254$$



Es $x = (1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 5, 5, 6, 6)$

$\bar{x} = 3 \quad , \quad var_e(x) = \frac{1}{14}\left(3 + 16 + 27 + 50 + 72 - 14 \cdot 9\right) =$

$x = (1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5) \qquad \bar{x} = 3$

$var_e(x) = 4 \quad , \quad \sigma_e(x) = 2$

OSS $F(a) = \sum_{i=1}^{n} (x_i - a)^2 \qquad \min F(a) = F(\bar{x})$

$$F'(a) = 2 \sum_{i=1}^{n} (x_i - a)(-1) = 0 \iff \sum_{i=1}^{n} (x_i - a) = 0$$

$$\iff \sum_{i=1}^{n} x_i - na = 0 \iff a = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

OSS $x = (x_1, \dots, x_n)$ e $y = (y_1, \dots, y_n)$ t.c.

$$y_i = b\, x_i + a \quad , \quad a, b \in \mathbb{R} .$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} (b x_i + a) = b\bar{x} + a$$

Def Varianza campionaria

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

Varianza empirica

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

OSS $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} \left(x_i^2 + \bar{x}^2 - 2 x_i \bar{x}\right) = \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i =$

$$= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}(n\bar{x}) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

<u>OSS</u> $\quad \mathrm{var}(x) = 0 \iff \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0 \iff (x_i - \bar{x})^2 = 0 \ \forall i$

$$\iff x_i = \bar{x} \quad \forall i$$

<u>Scarto quadratico medio campionario</u> $\quad \sigma(x) := \sqrt{\mathrm{var}(x)}$
(deviazione standard)

" $\quad$ <u>empirico</u> $\quad \sigma_e(x) := \sqrt{\mathrm{var}_e(x)}$

<u>Prop</u> (disuguaglianza di Chebyshev) Dati $x = (x_1, \dots, x_n)$, per ogni $d > 0$ si ha

$$\frac{\#\{x_i : |x_i - \bar{x}| > d\}}{n} \le \frac{\mathrm{var}_e(x)}{d^2} \quad \left(\le \frac{\mathrm{var}(x)}{d^2} \frac{n-1}{n} \right)$$

<u>dim</u> $\quad \sum_{i=1}^{n} (x_i - \bar{x})^2 \ge \sum_{\substack{i=1 \\ |x_i - \bar{x}| > d}}^{n} (x_i - \bar{x})^2 \ge \sum_{\substack{i=1 \\ |x_i - \bar{x}| > d}}^{n} d^2 =$

$$= d^2 \cdot \#\{x_i : |x_i - \bar{x}| > d\} \qquad \square$$

<u>ES</u> $\quad x = (1, 1, \dots)$

$$0.29 \sim \frac{4}{14} = \frac{\#\{x_i : |x_i - 3| > d\}}{14} \le \frac{1.571}{d^2} \quad \left(d = \frac{3}{2}\right) \sim 0.7$$

$$1 = \frac{\#\{x_i : |x_i - 3| > d\}}{14} \le \frac{4}{d^2} \quad \left(d = \frac{3}{2}\right) = \frac{16}{9}$$

<u>Cor</u> $\quad \frac{\#\{x_i : |x_i - \bar{x}| \le k\,\sigma(x)\}}{n} \ge 1 - \frac{1}{k^2} \quad \forall k \ge 1$

<u>dim</u> $\quad \#\{x_i : |x_i - \bar{x}| \le k\,\sigma(x)\} = n - \#\{x_i : |x_i - \bar{x}| > \overbrace{k\,\sigma(x)}^{d}\} \ge$

$$\geq n - n\left(\frac{var_e(x)}{k^2\sigma^2(x)}\right) = n - n\left(\frac{n-1}{n}\ \frac{var(x)}{k^2\sigma^2(x)}\right) =$$

$$= n - (n-1)\frac{\sigma^2(x)}{k^2\sigma^2(x)} = n - \frac{n-1}{k^2} = n\left(1 - \frac{n-1}{n}\frac{1}{k^2}\right) \geq$$

$$\geq n\left(1 - \frac{1}{k^2}\right) \qquad \square$$

<u>OSS</u> Per la distribuzione normale

perc. di dati in $[\bar{x} - \sigma(x), \bar{x} + \sigma(x)] \sim 68\%$

" " $[\bar{x} - 2\sigma(x), \bar{x} + 2\sigma(x)] \sim 95\%$

" " $[\bar{x} - 3\sigma(x), \bar{x} + 3\sigma(x)] \sim 99.7\%$

<u>OSS</u> $y_i = bx_i + a$, $var(y) = b^2\, var(x)$

$\sigma(y) = b\,\sigma(x)$