

Lezione 8: Statistica

[Dati perfetti]

PROBABILITÀ

???

STATISTICA

[Dati imperfetti]

modellizzazione
complessa

Caos
quantità di dati

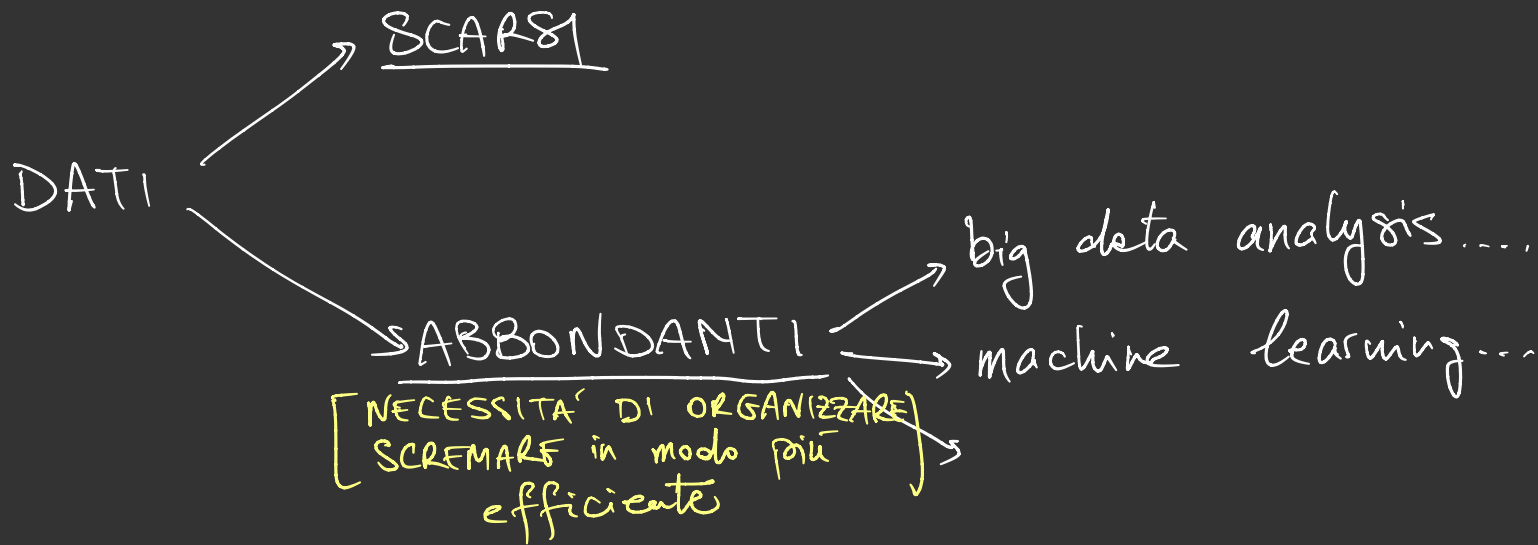
MODELLO

FREQUENZE
relative / assolute

misurazioni

"nel lungo periodo le
misurazioni/osservazioni
convergeranno al modello
perfetto"

LA LEGGE DEI GRANDI NUMERI



Partiamo da un set di dati.

PESO DI 15 CAVIE: { 28, 32, 37, 29, 31, 30, 32, 26, 32, 27 }
(in g)

37 is highlighted as an **OUTLIER**.

Obiettivo: trovare modi per descrivere facilmente un set di dati

Voglio un \bar{x} tale che $\bar{x} - x_i$ siano piccole.

minimizzare $\sum_{i=1}^{n=15} (\bar{x} - x_i) = \underline{n \cdot \bar{x} - \sum x_i = 0}$

↑ impiego.
SOMMA DEGLI ERRORI = 0

$\Rightarrow \bar{x} = \frac{\sum x_i}{n}$ MEDIA ARITMETICA.

Qualcuno potrebbe obiettare: non è giusto prendere gli errori con segno positivo o negativo.

$f(x) := \sum_{i=1}^m \frac{(x - x_i)^2}{\text{SCARTO QUADRATICO}}$

↑ definisco una funzione

in questo modo gli errori sono sempre positivi e non hanno speranze di semplificarsi.

Allora saremo interessati al MINIMO di $f(x)$.

$$f(x) = \sum_i (x - x_i)^2 = \sum_i (x^2 - 2xx_i + x_i^2) = \underbrace{n}_{a} x^2 - 2(\underbrace{\sum x_i}_{b})x + \underbrace{\sum x_i^2}_{c}$$

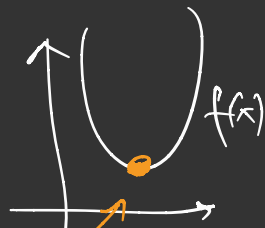
Questa è una funzione quadratica in x .

$$f(x) = ax^2 + bx + c$$

Guardiamo se $a > 0$ oppure $a < 0$. $a = n > 0 \Rightarrow$

minimo $f(x) = -\frac{b}{2a} = \frac{\sum x_i}{n} = \text{MEDIA ARITMETICA.}$

alora $\exists!$ MINIMO!



• $\sum (\bar{x} - x_i) = 0$

• minimizzare $\sum (x - x_i)^2$ } \rightarrow dà lo stesso risultato!
[la media].

E se avessimo scelto $\sum |x - x_i|$??

avremmo ottenuto quello che si chiama mediana.

Ordinate il set di dati:

$$26 \leq 27 \leq 28 \leq 28 \leq 29 \leq 29 \leq 30 \leq \boxed{30} \leq 31 \leq 31 \leq 31 \leq 32 \leq 32$$

$\overset{x_1}{\underset{''}{26}}$ $\overset{x_2}{\underset{''}{27}}$ $\overset{x_8}{\underset{''}{30}}$ $\leq 32 \leq 37 \overset{x_{15}}{\underset{''}{}}$

Mediana $M =$ dato centrale

\downarrow

$= \begin{cases} X_{\frac{(n+1)}{2}} & n \text{ dispari} \\ \frac{(X_{\frac{n}{2}} + X_{\frac{n}{2}+1})}{2} & n \text{ pari} \end{cases}$

In questo caso

$$M = X_{\frac{15+1}{2}} = 8 = X_8$$

$$\downarrow$$

$$= 30$$

Se buttiamo via 37

$$M = \frac{30 + 30}{2} = 30$$

In generale: quando abbiamo un set di dati finito possiamo metterlo in ordine:

MEDIANA

112

secondo QUARTILE

terzo QUARTILE

$X_1 \leq X_2 \leq \dots \leq X_s \leq \dots \leq X_{1000000} \leq \dots \leq X_n$

primo PERCENTILE.

primo QUARTILE
è quel valore che ha a sinistra il 25% dei dati

DEF 1 (VARIANZA) (SCARTO QUADRATICO MEDIO)

$$\text{Var}(x_i) := \frac{\sum_{i=1}^n (\bar{x} - x_i)^2}{n} \quad (\bar{x} = \text{MEDIA degli } x_i)$$

(DEVIAZIONE STANDARD) $\sigma(x_i) := \sqrt{\text{Var}(x_i)} = \sqrt{\frac{\sum (\bar{x} - x_i)^2}{n}}$

(COEFFICIENTE di VARIAZIONE) $\text{CV}(x_i) := \frac{\sigma(x_i)}{\bar{x}}$

OSS 1 Se $x_i = x_j \quad \forall (i, j) \Rightarrow \begin{cases} \text{Var}(x_i) = \sigma(x_i) = 0 \\ \bar{x} = x_i \quad \forall i = 1, \dots, n. \end{cases}$

OSS 2 $\text{Var}(x_i) = \frac{1}{n} \sum_{i=1}^n (\bar{x} - x_i)^2 = \frac{1}{n} \sum (\bar{x}^2 - 2\bar{x}x_i + x_i^2)$

$$= \frac{1}{n} \cdot \bar{x}^2 \cdot \underbrace{\sum_{i=1}^n 1}_{=n} - \frac{2}{n} \bar{x} \left(\sum_{i=1}^n x_i \right) + \frac{\sum x_i^2}{n}$$
$$= \bar{x}^2 - 2\bar{x}^2 + \frac{\sum x_i^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2 = \underline{\text{Media}(x_i^2) - \text{Media}(x_i)^2}$$

$$\underline{\text{Var}(x_i) = \text{Media}(x_i^2) - (\text{Media}(x_i))^2}$$

Gestire i debiti è complesso. Bisogna fare i conti con l'incertezza \rightarrow si fanno delle stime.

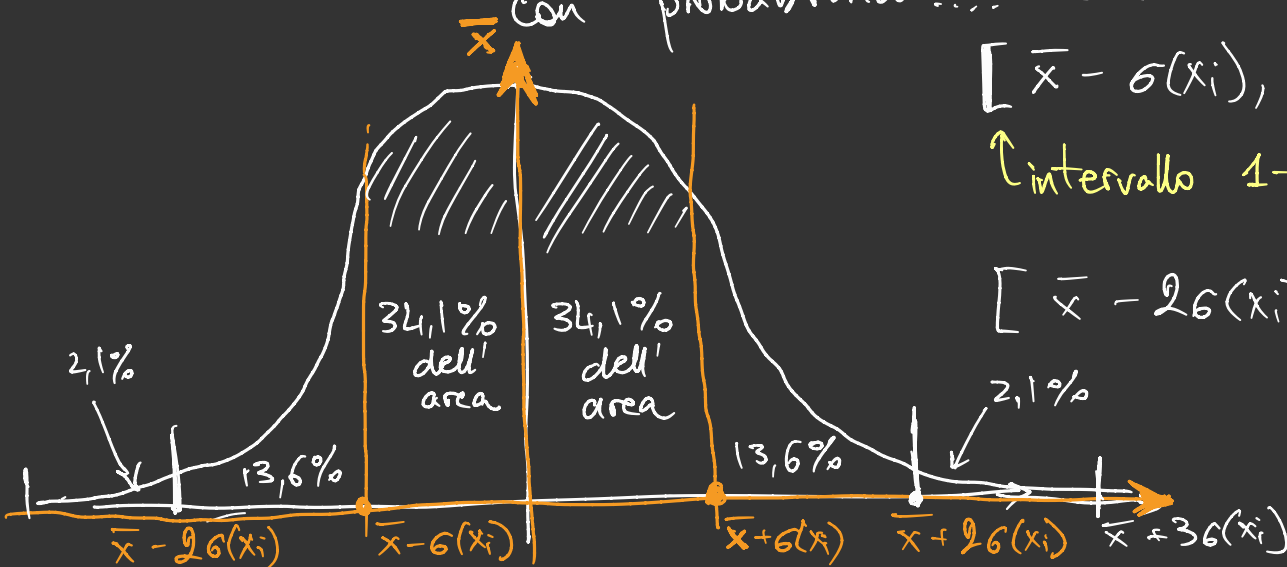
Per esempio: il risultato che cerchiamo con probabilità... sta dentro all'intervallo

$$[\bar{x} - \sigma(x_i), \bar{x} + \sigma(x_i)]$$

\uparrow intervallo 1-sigma.

$$[\bar{x} - 2\sigma(x_i), \bar{x} + 2\sigma(x_i)]$$

2-sigma.



Es: nella distribuzione normale (che vedremo).

| | | | | | |
|-----------|--|-------|-----------|------------|------|
| 1 - sigma | $[\bar{x} - \sigma, \bar{x} + \sigma]$ | ha il | 68,3% | dell' area | |
| 2 - sigma | $[\bar{x} - 2\sigma, \bar{x} + 2\sigma]$ | " | 95,5% | " | ← 2σ |
| 3 - sigma | $[\bar{x} - 3\sigma, \bar{x} + 3\sigma]$ | " | 99,7% | " | " |
| 4 - sigma | $[\bar{x} - 4\sigma, \bar{x} + 4\sigma]$ | " | 99,993% | " | " |
| 5 - sigma | $[\bar{x} - 5\sigma, \bar{x} + 5\sigma]$ | " | 99,99994% | " | " |

BOSONE DI HIGGS
← 5σ

Sono misure dell' incertezza.

Di solito in statistica ci si accontenta del 2-sigma.

§. RETTA DI REGRESSIONE:

Supponiamo di avere 2 sets di dati.

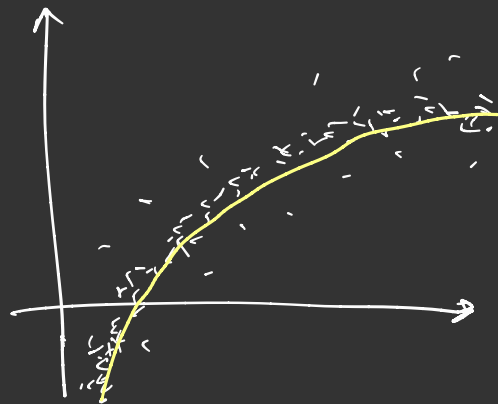
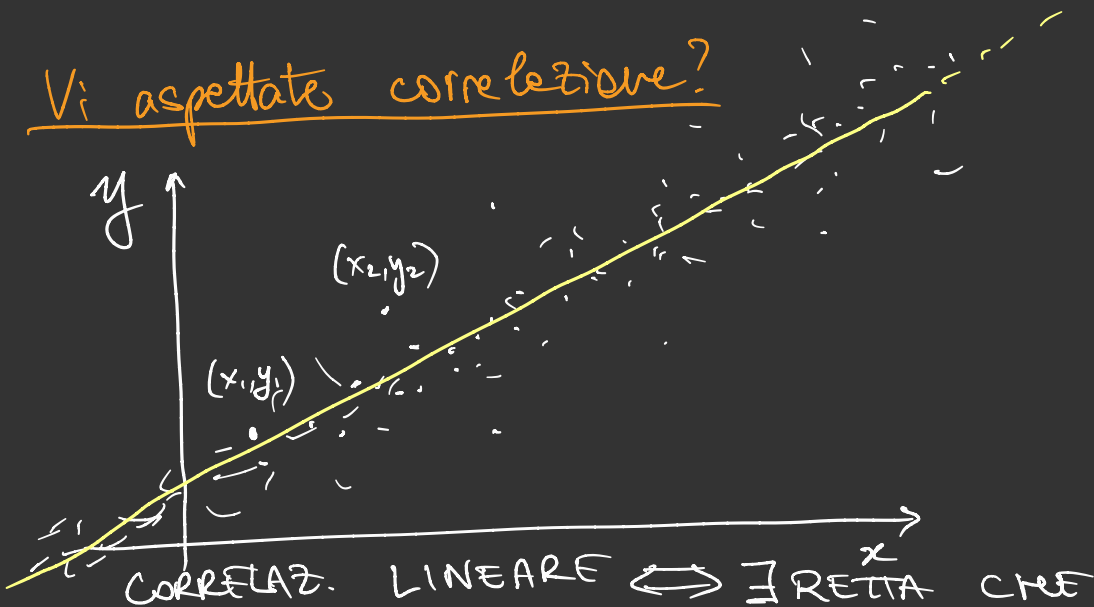
| | | | | | | |
|-------|-------|-------|-----|-----|-----|-------|
| x_1 | x_2 | x_3 | ... | ... | ... | x_n |
| y_1 | y_2 | y_3 | ... | ... | ... | y_n |

Esempio:

x_n ← ml di pioggia caduti nell'anno i

y_n ← chili di pannocchie cresciute nell'anno i .

Vi aspettate correlazione?

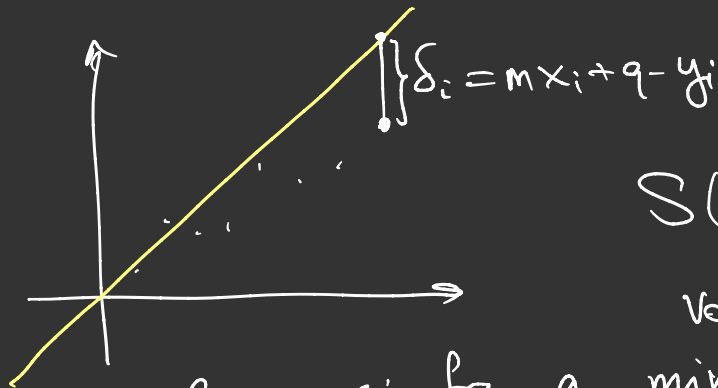


DESCRIVE LA CORRELAZIONE TRA le x_i e le y_i .

$$y = \underline{m}x + \underline{q}$$

1. Basterebbe trovare $\begin{cases} m(x_i, y_i) \\ q(x_i, y_i) \end{cases}$

2. Ci vorrebbe un modo per sapere quanto questa retta approssime bene la correlazione.



$$S(m, q) := \frac{1}{n} \sum_{i=1}^n (mx_i + q - y_i)^2$$

Voglio minimizzare questo.

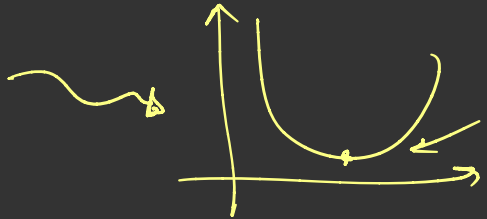
Come si fa a minimizzare una funzione in 2 variabili?

• Prima una e poi l'altra!

Prima guarda a $q \mapsto S(m, q)$

$$S(m, q) = \left(\frac{1}{n}\right) \sum_{i=1}^m (x_i^2 m^2 + q^2 + y_i^2 + 2x_i m q - 2x_i y_i m - 2y_i q)$$

$$\downarrow = q^2 \cdot \underbrace{1}_{=a} + q^1 \cdot \underbrace{2(m\bar{x} - \bar{y})}_{b} + q^0 \left[\frac{m^2}{n} \sum x_i^2 - \frac{2m}{n} \sum x_i y_i + \frac{1}{n} \sum y_i^2 \right]$$

$a=1 > 0$  quindi $\exists!$ MINIMO:
lo chiamo $q_0(m)$.

$$q_0(m) = -\frac{b}{2a} = \bar{y} - m\bar{x}$$

$$\Rightarrow S(m, \underbrace{q_0(m)}_{\substack{\uparrow \\ \text{MINIMO per } q}}}) = c - \frac{b^2}{4a} = m^2 \underbrace{\left[\frac{\sum x_i^2}{n} - \bar{x}^2 \right]}_{\substack{= \text{Var}(x_i) > 0 \\ \tilde{a}}} - 2 \underbrace{\left[\frac{\sum x_i y_i}{n} - \bar{x} \bar{y} \right]}_{\tilde{b}} m + m^0 \underbrace{\left[\frac{\sum y_i^2}{n} - \bar{y}^2 \right]}_{\tilde{c}}$$

$\Rightarrow S \tilde{c}$  anche come funzione \tilde{c} di $m!$

$\Rightarrow \exists!$ minimo per $m \mapsto S(m, q_0(m))$

$$\text{il minimo è } \bar{m} = -\frac{b}{2a} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2}$$

ABBIAMO OTTENUTO LA NOSTRA RETTA DI REGRESSIONE:

$$y = (\bar{m})x + (\bar{q}) \quad \left\{ \begin{array}{l} \bar{m} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \quad \text{es.} \quad \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ \bar{q} = \bar{y} - \bar{m} \cdot \bar{x} \end{array} \right.$$

Quanto è buona questa retta?

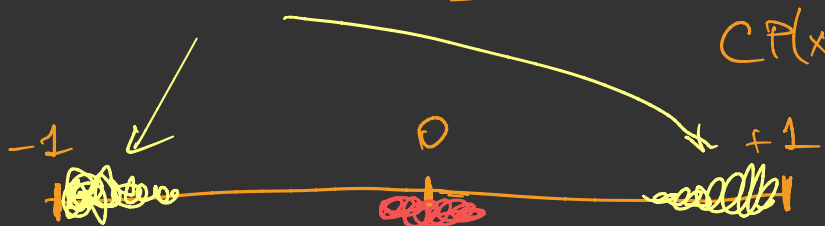
$S(\bar{m}, \bar{q})$ è assoluto, lo voglio relativizzare
[cioè CONFRONTARE con qualcosa]

La quantità relativa e^- :

$$\frac{S(\bar{m}, \bar{q})}{\text{Var}(y_i)} = 1 - \frac{\overbrace{(\overline{xy} - \bar{x} \cdot \bar{y})^2}^{>0}}{\underbrace{\text{Var}(x_i)}_{>0} \underbrace{\text{Var}(y_i)}_{>0}} \in \underline{[0, 1]}$$

$$e^- \text{ zero} \Leftrightarrow \frac{(\overline{xy} - \bar{x} \cdot \bar{y})^2}{\text{Var}(x_i) \text{Var}(y_i)} = 1 \quad \Leftrightarrow \sqrt{\cdot} \quad \boxed{\frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma(x_i) \sigma(y_i)} = \pm 1.}$$

La retta descrive bene i dati $[-1, -0,95] \cup [0,95, 1]$



$$CP(x, y) =$$

COEFFICIENTE DI PEARSON DELLA RETTA DI REGRESSIONE

$$y = \bar{m} \cdot x + \bar{q}$$

LA RETTA NON DESCRIVE BENE I DATI