MATEpristem

Giochi matematici e non solo: sfide e parole–chiave

Roma, 29 settembre – 1 ottobre 2017

Incertezza delle misure e misura dell'incertezza: un percorso tra Statistica e Probabilità

Walter Racugno
Università di Cagliari

Premesse

- Incertezza!!!
- Non è una lezione
- Pre-conoscenze sulle basi della Statistica e della Probabilità
- Qualche richiamo (intuitivo)
- Un percorso tra Statistica e Probabilità (con qualche riflessione)
- Iniziamo con alcune precisazioni sulla Statistica

Di che cosa si occupa la Statistica?

- La Fisica di fenomeni *naturali*
- Sociologia: fenomeni *sociali*
- Geologia: fenomeni che riguardano la *crosta terrestre*
- Astronomia: fenomeni *celesti*
- Biologia: fenomeni della vita (*biologici*)
- Medicina: fenomeni che riguardano lo *stato di salute*
- Economia: fenomeni di *gestione delle risorse*
- Chimica: fenomeni sulla composizione e trasformazioni della *materia*
- . . . . . .

La Statistica si occupa di ***fenomeni reali*** !

Si “presta” dunque a tutte le altre discipline.

affermazione un po’ spocchiosa ... ma è anche vero che lo statistico non si sostituisce mai all’esperto di dominio!

Di che cosa si occupa la Statistica?

Per studiare un fenomeno (reale) è necessario, in una prima fase, acquisire informazioni su di esso:

- Osservazioni sperimentali (rilevazioni dei dati)
- Raccolta, organizzazione e sintesi dei dati (tabelle, grafici, indici)
- Prime interpretazioni del fenomeno e formulazione di ipotesi
- Seconda fase: inferenza statistica

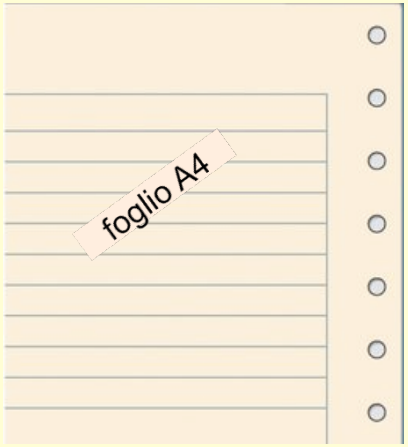
In genere, siamo abituati ad associare la Statistica allo studio di fenomeni collettivi: fenomeni a cui concorrono una molteplicità di soggett (individui, unità statistiche), tutti aventi il medesimo carattere, o caratteri, d'interesse.

Il carattere - qualitativo ordinabile o sconnesso, oppure quantitativo discreto o continuo - si manifesta con diverse modalità (o determinazioni), che sono l'oggetto delle rilevazioni.

L'insieme delle unità statistiche è omogeneo rispetto ai caratteri in studio.

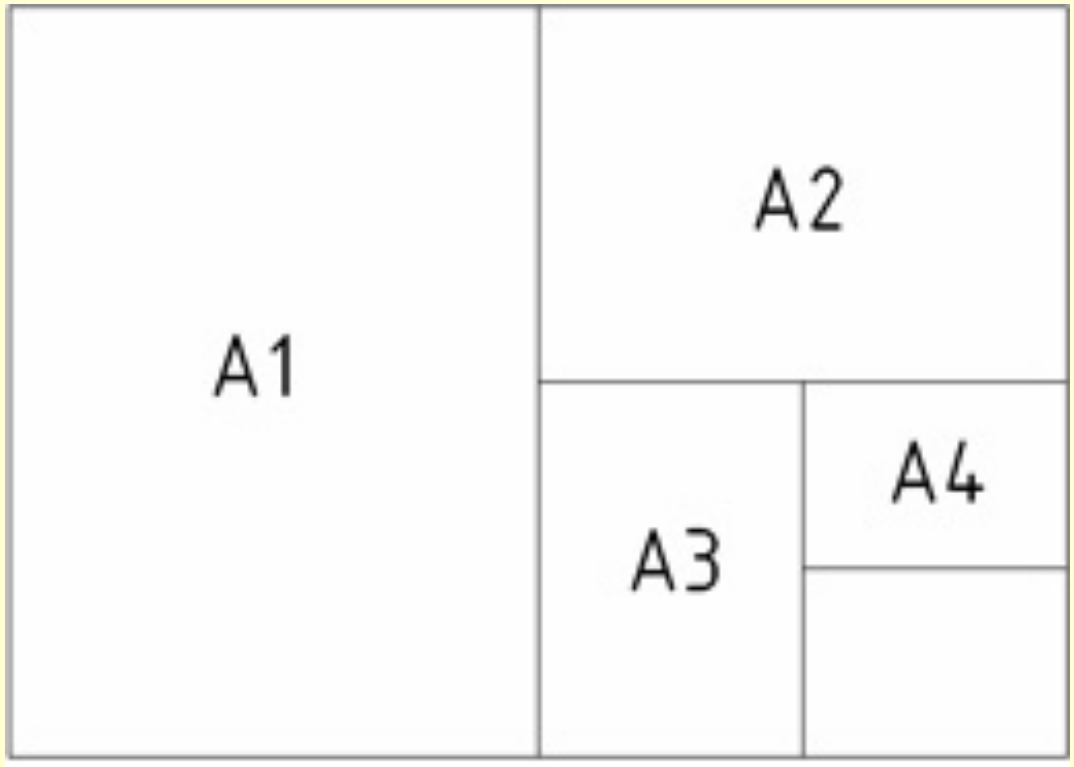
Incertezza delle misure e misura dell'incertezza

vogliamo misurare la lunghezza di un foglio A4



una breve digressione

Formato A4 (*21.0 x 29.7 cm)*



A0 (841x1189 mm) =1 m^2

tagliando a metà ogni metà, per 4 volte ⟶ A4 (210x297)

I lati ℓ e L di ciascun formato sono tutti in rapporto $\sqrt{2}$ tra loro:

$$L = \ell \times \sqrt{2}$$

Perché $\sqrt{2}$

Il formato A è stato definito considerando un foglio di area = 1 m^2 foglio (A0), con lati ℓ ed L tali che dimezzandolo si avesse un nuovo foglio con lati aventi ancora le medesime proporzioni.

E così per ogni successivo dimezzamento.

Nota: in questo modo, qualsiasi formato può – ovviamente – essere usato per costruire una tassellazione dei formati più grandi o, in altri termini, i formati piccoli (sotto-insiemi) producono una partizione di qualsiasi formato (insieme) più grande. Si ha dunque un inseme di sottoinsiemi necessari e sufficienti, tra loro proporzionali.

Deve quindi essere $\frac{L}{\ell} = \frac{\ell}{L/2} \rightarrow L = \ell\sqrt{2}$

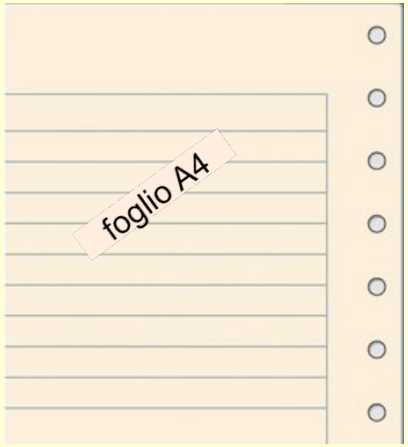
imponendo $\ell \times L = 1\, m^2$

si ha: $L = \sqrt[4]{2} \approx 1.189\, m$ e quindi $\rightarrow$ dimensioni di A0 : $841 \times 1189\, mm$ (circa)

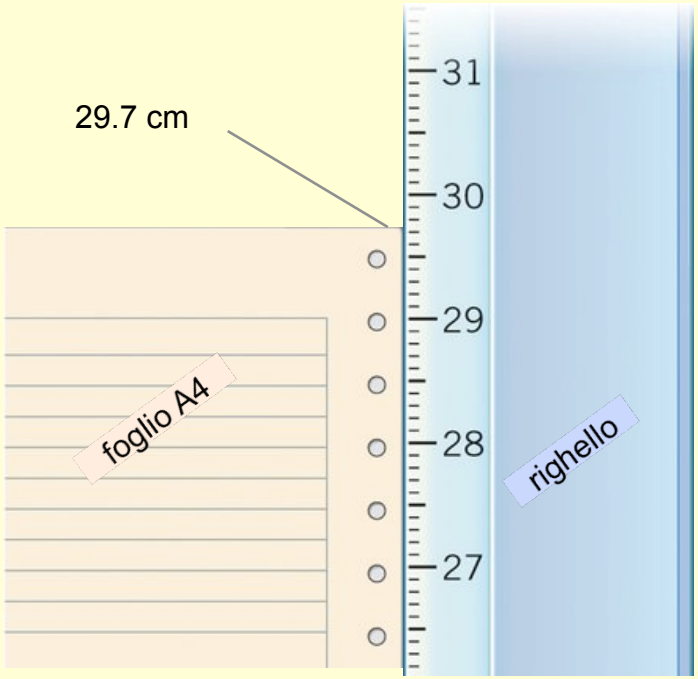
dimezzando 4 volte $\rightarrow$ A4 : $210 \times 297\, mm$

fine digressione

vogliamo misurare la lunghezza di un foglio A4



lunghezza di un foglio A4

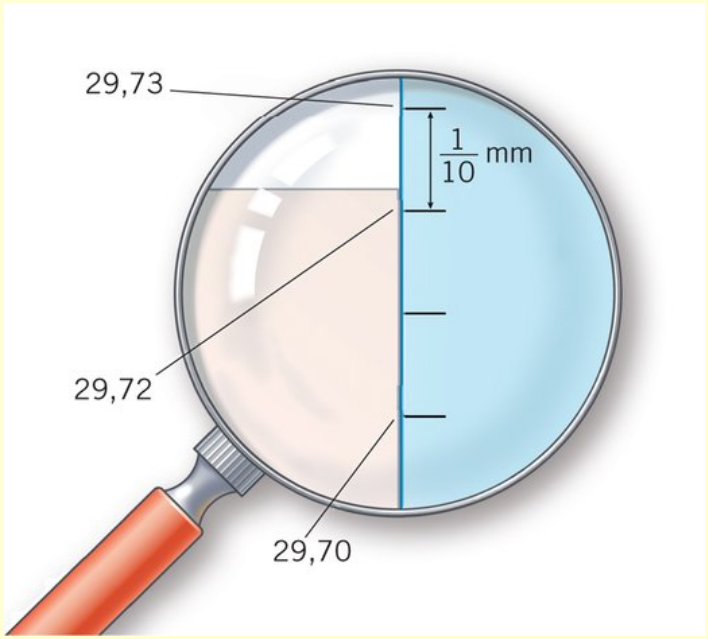


con un righello che ha la *sensibilità* di 1 *mm* ...

... ma ...

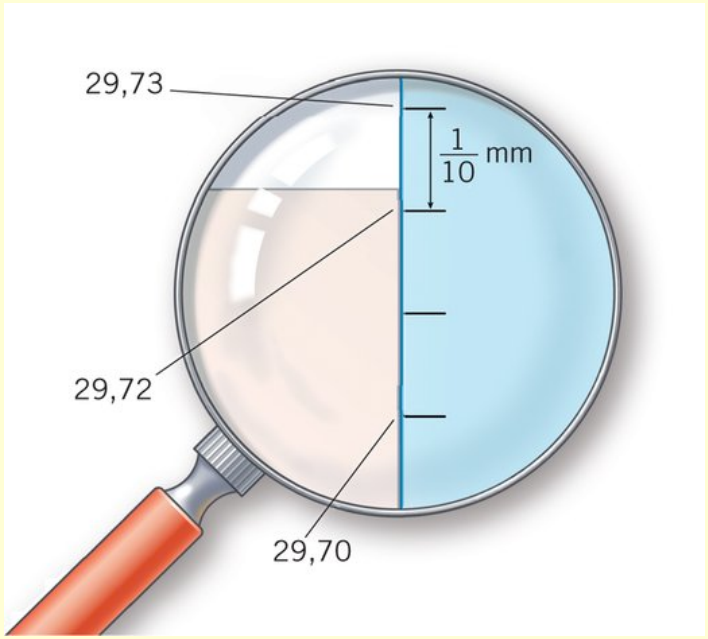
volendo essere un poco più precisi ...

al decimo di millimetro (*dmm*)



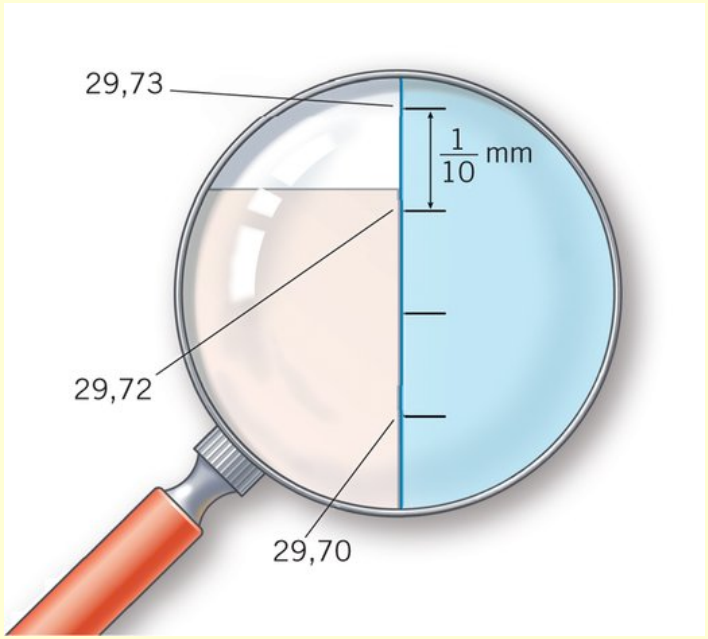
29.72

al decimo di millimetro (*dmm*)



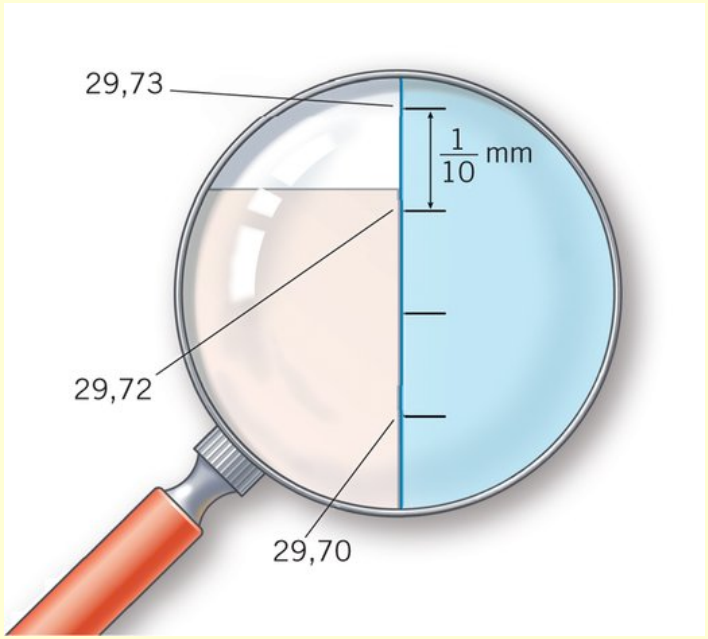
29.72 29.73

al decimo di millimetro (*dmm*)



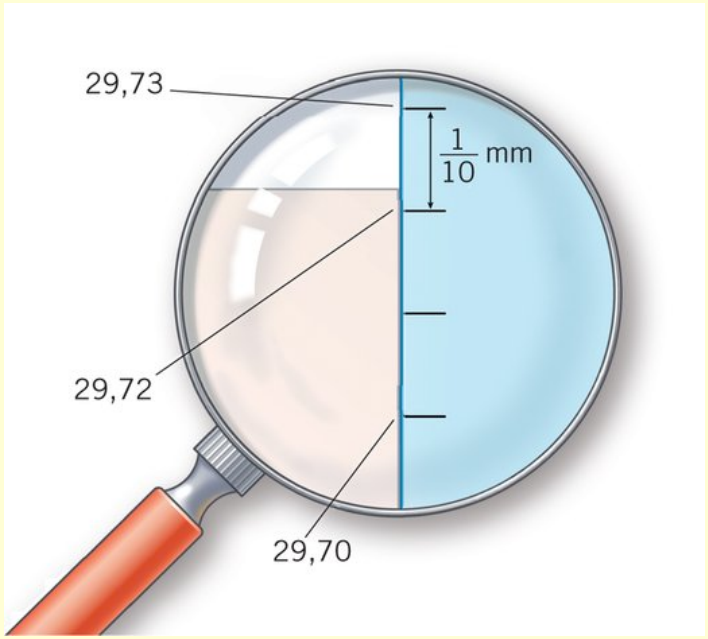
29.72 29.73 29.71 29.70

al decimo di millimetro (*dmm*)



29.72 29.73 29.71 29.70 29.72 29.72 29.73 29.76 29.72 29.70

al decimo di millimetro (*dmm*)



29.72	29.73	29.71	29.70	29.72	29.72	29.73	29.76	29.72	29.70
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}

più sinteticamente: y_i con i =1, 2, ..., n (nel nostro caso n = 10)

---

figure da: http://ebook.scuola.zanichelli.it/amalditraiettorie/

... e se usassimo il centesimo di millimetro (*cmm*) ?

o il millesimo di *mm* (*micron* μ), cioè il milionesimo di metro: $1\mu m=10^{-6}\ m$?

finiremmo per entrare nella struttura cellulosica della carta e ci accorgeremmo che il bordo del foglio è sfilacciato dalle fibre della cellulosa con conseguente aumento dell'incertezza, certamente molto ridotta rispetto al millimetro o al centesimo di millimetro, ma con una variabilità molto elevata tra le diverse rilevazioni (al livello dei *micron*).

E se andassimo a misurare le dimensioni di un diamante? che non ha "sfilacciature" o fibre sulla sua superficie.

Con strumenti sempre più sensibili, entreremmo nella struttura molecolare del cristallo e l'oscillazione delle particelle elementari ci impedirebbe di stabilire la misura esatta (*principio di indeterminazione di Heisenberg*).

E’ impossibile ottenere misure esatte:

l’incertezza di una misura può essere ridotta ma mai eliminata!

La si può misurare

NOTA

errori casuali (accidentali, statistici, aleatori): non controllabili, non eliminabili, dovuti a fattori intrinseci (natura dell’oggetto misurato) e esterni (strumento di misura/misuratore, fluttuazioni ambientali, ecc.). Possono alterare la misura sia in eccesso sia in difetto.

errori sistematici: sono costanti in entità e mantengono lo stesso verso (in eccesso o in difetto) → errori eliminabili.

lunghezza di un foglio A4

29.72	29.73	29.71	29.70	29.72	29.72	29.73	29.76	29.72	29.70
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}

I dati ordinati

29.70	29.70	29.71	29.72	29.72	29.72	29.72	29.73	29.73	29.76
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

(x_i con i =1, 2, ..., n)

Una prima valutazione dell'incertezza dovuta agli errori casuali si ha calcolando il *range* (intervallo di variazione) delle misure osservate

$$range = x_{max} - x_{min}$$ (nel nostro caso: 29.76 – 29.70 = 0.06 *cm*)

La *semidispersione* (o *errore massimo*): $\varepsilon = range/2$ fornisce dunque l'errore massimo per eccesso o per difetto che si può essere commesso nelle n misurazioni.

Il risultato della misura, cioè la lunghezza del foglio A4, sarà un valore compreso nell'intervallo:

$$(\ell - e\ ,\ \ell + e)$$

Quale valore attribuire a ℓ ?

Poiché gli errori casuali si verificano sia in eccesso sia in difetto, rispetto al valore ℓ, e senza alcuna sistematicità, il valore più plausibile per ℓ è il valore medio delle *n* misure osservate, cioè quel valore che rende nulla la somma degli scarti delle diverse misure da esso (giacché gli scarti positivi e negativi si compensano):

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \bar{x}$$

o, più semplicemente, $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$

lunghezza di un foglio A4

Nell'esempio

29.70	29.70	29.71	29.72	29.72	29.72	29.72	29.73	29.73	29.76
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

(x_i con i =1, 2, ..., 10)

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{297.21}{10} = 29.721\ cm$$

quindi

$$lunghezza\ foglio = \bar{x} \pm e = 29.721 \pm 0.06\ (cm)$$

lunghezza di un foglio A4

Nell'esempio

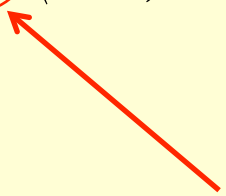
29.70	29.70	29.71	29.72	29.72	29.72	29.72	29.73	29.73	29.76
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

(x_i con i =1, 2, ..., 10)

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{297.21}{10} = 29.721\ cm$$

quindi

$$lunghezza\ foglio = \bar{x} \pm e = 29.721 \pm 0.06\ (cm)$$



Alcune sintesi dei dati osservati

29.70	29.70	29.71	29.72	29.72	29.72	29.72	29.73	29.73	29.76
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

raggruppando le misure uguali

x_i	n_i
29.70	2
29.71	1
29.72	4
29.73	2
29.76	1
	$n = 10$

$i = 1,\ 2,\ ...,\ k \quad (k \le n)$

$$\sum_{i=1}^{k} n_i = n$$

nell'esempio: $i = 1,\ 2,\ 3,\ 4,\ 5$

$$\sum_{i=1}^{5} n_i = 10$$

n_i molteplicità (o *frequenza*) della misura x_i

alcune sintesi dei dati osservati

Se le misurazioni fossero state *n*=100? (*n=1*000? *n*=10000?)

Immaginando che nella tabella le proporzioni siano le stesse!

x_i	n_i
29.70	20
29.71	10
29.72	40
29.73	20
29.76	10
	$n = 100$

$$i = 1,\ 2,\ 3,\ 4,\ 5$$

$$\sum_{i=1}^{5} n_i = 100$$

alcune sintesi dei dati osservati

Se le misurazioni fossero state *n*=1000?

Immaginando che nella tabella le proporzioni siano le stesse!

x_i	n_i
29.70	200
29.71	100
29.72	400
29.73	200
29.76	100
	$n = 1000$

$$i = 1,\ 2,\ 3,\ 4,\ 5$$

$$\sum_{i=1}^{5} n_i = 1000$$

alcune sintesi dei dati osservati

Per liberarci da *n* (*fattore di disturbo*) e consentire il confronto tra studi con diverse numerosità di misurazioni: $f_i = \frac{n_i}{n}$ frequenza con cui si presenta la stessa misura x_i rispetto al totale delle *n* misurazioni effettuate

x_i	n_i	$f_i = \frac{n_i}{n}$
29.70	2	0.2
29.71	1	0.1
29.72	4	0.4
29.73	2	0.2
29.76	1	0.1

$n = 10$

f_i *frequenza relativa*:

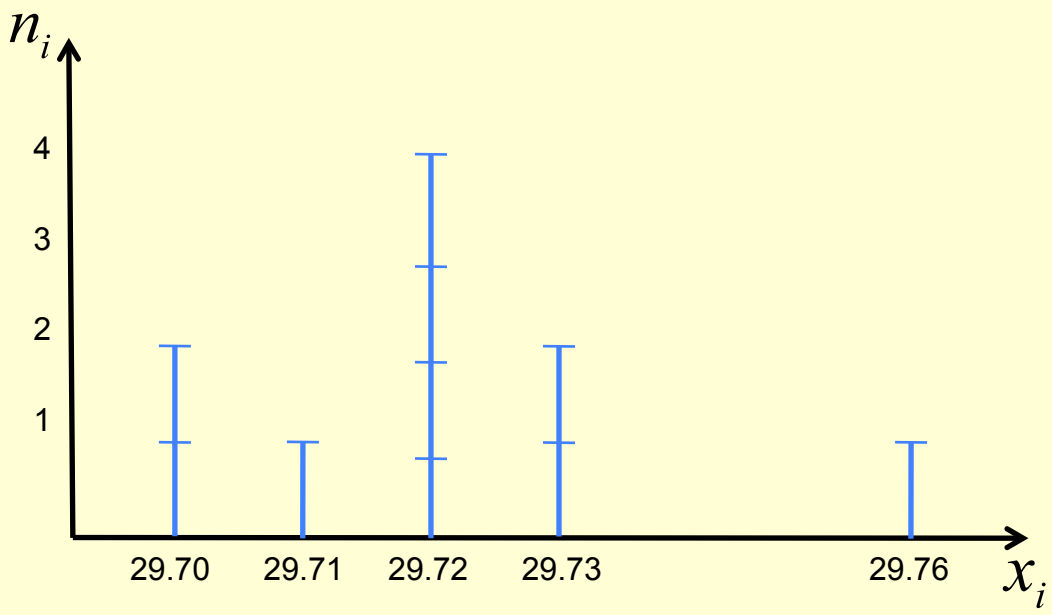
"peso" della rilevazione x_i rispetto alla totalità delle *n* rilevazioni

$$\sum_{i=1}^{5} f_i = 1$$

Una rappresentazione grafica: *diagramma a barre*

Dove la lunghezza di ciascuna barra rappresenta la frequenza n_i di ciascuna misura x_i

x_i	n_i
29.70	2
29.71	1
29.72	4
29.73	2
29.76	1
	$n = 10$

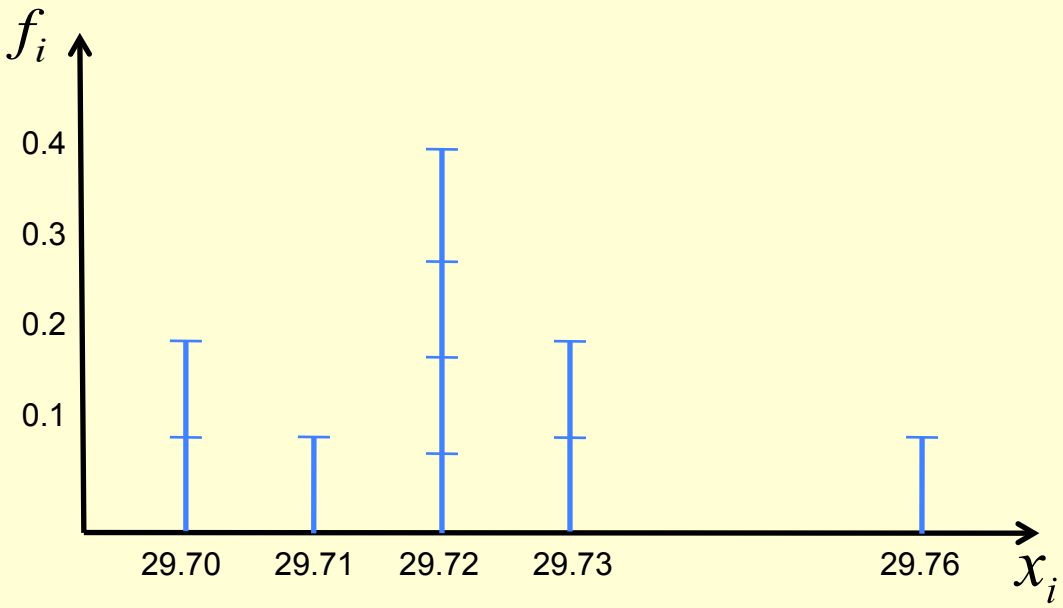


Analogamente per f_i

Dove la lunghezza di ciascuna barra rappresenta il “peso” f_i

(peso relativo = *frequenza relativa)* di x_i

x_i	f_i
29.70	0.2
29.71	0.1
29.72	0.4
29.73	0.2
29.76	0.1
	1

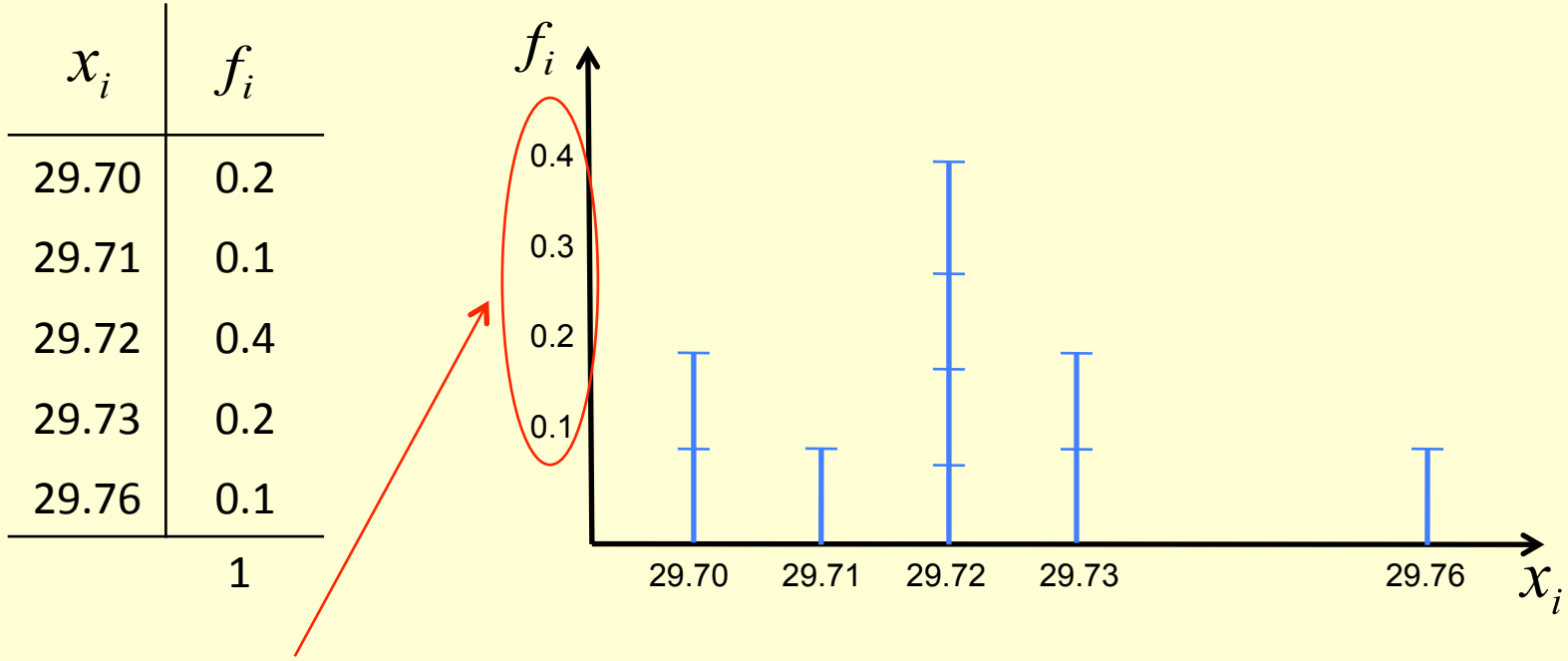


Analogamente per f_i

Dove la lunghezza di ciascuna barra rappresenta il “peso” f_i

(peso relativo = *frequenza relativa)* di x_i

x_i	f_i
29.70	0.2
29.71	0.1
29.72	0.4
29.73	0.2
29.76	0.1
	1



- dati
- frequenze assolute
- frequenze relative
- diagramma a barre
- . . . .

- dati
- frequenze assolute
- frequenze relative
- diagramma a barre
- . . . .

questa è Statistica!

allora usiamo le nozioni e le tecniche di questa disciplina, in particolare:

- la media $\overline{x}$
- la varianza σ^2 e, conseguentemente, la deviazione standard $sd = \sigma$ come misura della dispersione (*misura dell'incertezza*!)

Avvertenza

La Statistica studia *fenomeni collettivi*, le misurazioni del foglio A4 sono effettuate tutte sul medesimo oggetto, come possiamo trattare questo "fenomeno" come collettivo?

Possiamo assumere che:

a) le *n* misure siano rilevate su *n* fogli estratti casualmente da un risma,

oppure che

b) le misure siano eseguite sullo stesso foglio da *n* individui diversi,

nel caso *a)* le unità statistiche sono gli *n* fogli, caratterizzati dalla variabile lunghezza;

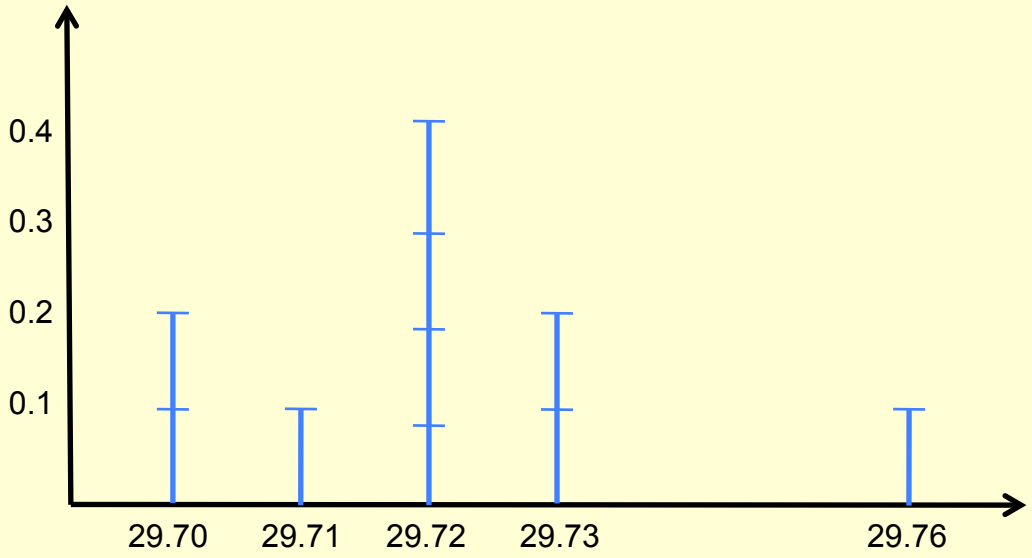
nel caso *b)* le unità statistiche sono gli *n* individui, caratterizzati dalla misura da loro osservata.

Più semplicemente*: le n misure sono tra loro indipendenti!*

Indici sintetici di una distribuzione di frequenza

x_i	n_i	$f_i = \frac{n_i}{n}$
29.70	2	0.2
29.71	1	0.1
29.72	4	0.4
29.73	2	0.2
29.76	1	0.1

$n = 10$

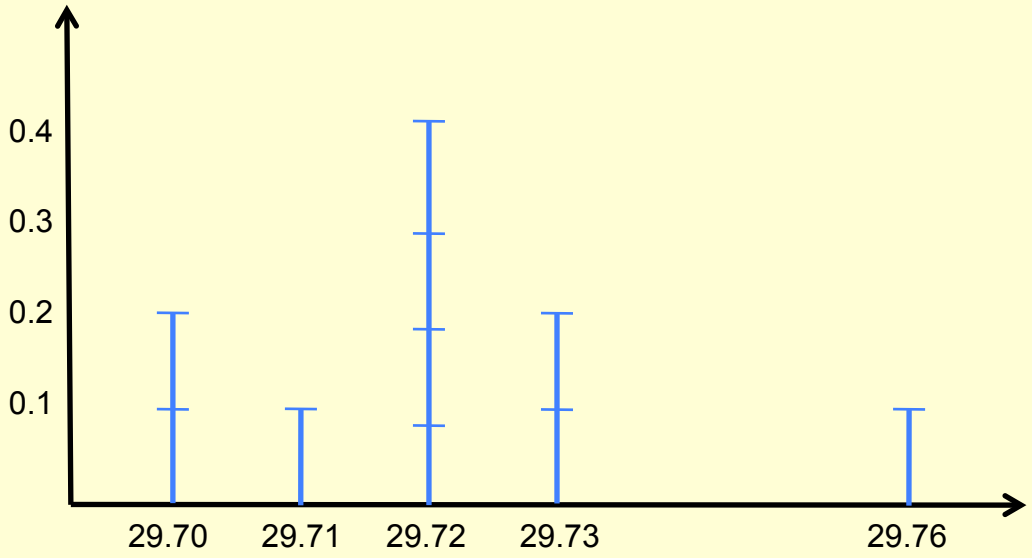


$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}\sum_{i=1}^{5} x_i n_i = \sum_{i=1}^{5} x_i f_i = 29.721\ cm$$

Indici sintetici di una distribuzione di frequenza

x_i	n_i	$f_i = \frac{n_i}{n}$
29.70	2	0.2
29.71	1	0.1
29.72	4	0.4
29.73	2	0.2
29.76	1	0.1

$n = 10$



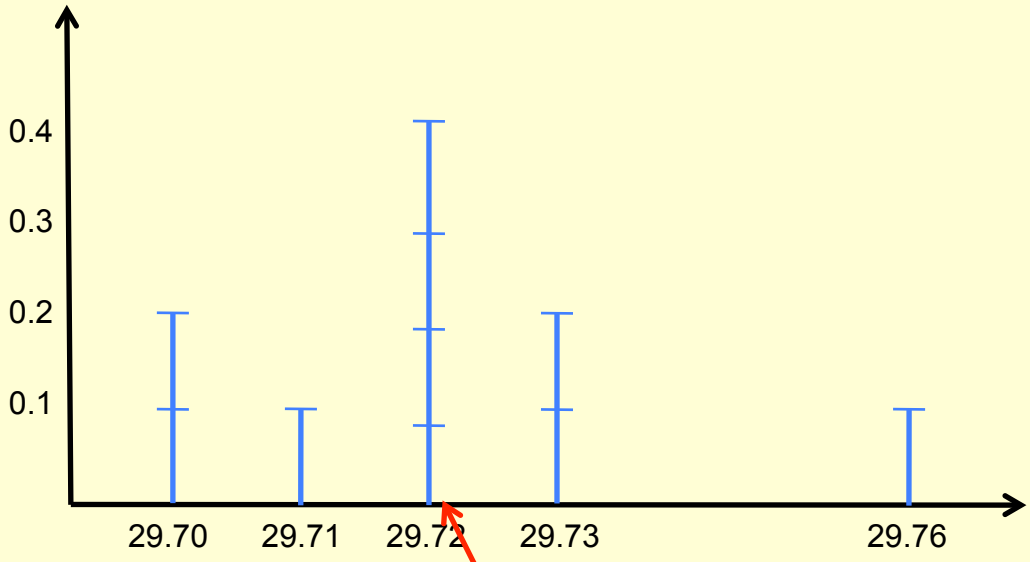
$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}\sum_{i=1}^{5} x_i n_i = \sum_{i=1}^{5} x_i f_i = 29.721\ cm$$

Valore delle x_i se fossero tutte uguali e a parità della loro somma $\sum_{i=1}^{10} x_i$

Indici sintetici di una distribuzione di frequenza

x_i	n_i	$f_i = \frac{n_i}{n}$
29.70	2	0.2
29.71	1	0.1
29.72	4	0.4
29.73	2	0.2
29.76	1	0.1

$n = 10$

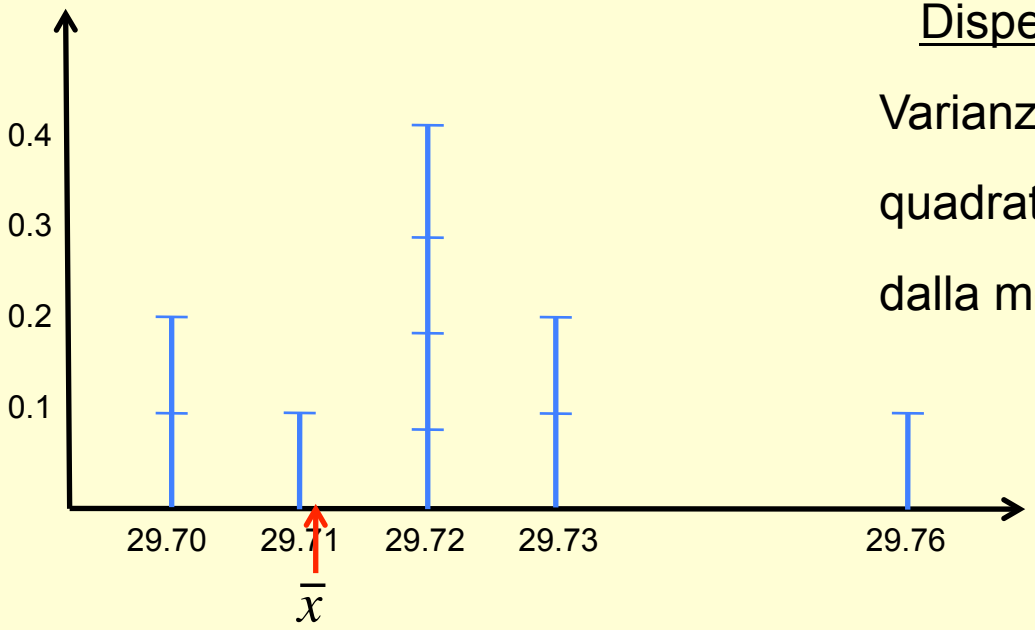


$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}\sum_{i=1}^{5} x_i n_i = \sum_{i=1}^{5} x_i f_i = 29.721\ cm$$

Valore delle x_i se fossero tutte uguali e a parità della loro somma $\sum_{i=1}^{10} x_i$

Baricentro della distribuzione dei “pesi”

Indici sintetici di una distribuzione di frequenza



<u>Dispersione (misura dell'incertezza)</u>

Varianza = media degli scarti $(x_i - \bar{x})$

quadratici $(x_i - \bar{x})^2$ delle misure x_i

dalla media $\bar{x}$

$$\sigma^2 = \frac{1}{10}\sum_{i=1}^{5}(x_i - \bar{x})^2 f_i = 0.00027\ cm^2 \rightarrow sd = 0.016\ cm$$

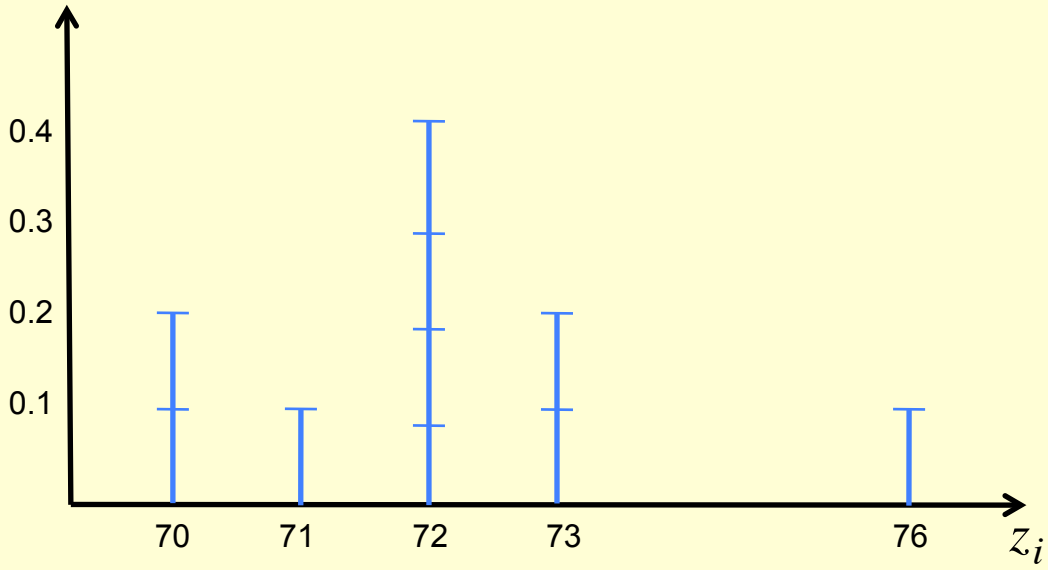
$$lunghezza\ foglio = 29.721 \pm 0.016\ cm$$

un’altra digressione

Il 29 si ripete sempre: perché perdere tempo a scriverlo? ~~29~~ (*cm*)

z_i	$f_i = \frac{n_i}{n}$
70	0.2
71	0.1
72	0.4
73	0.2
76	0.1

dmm



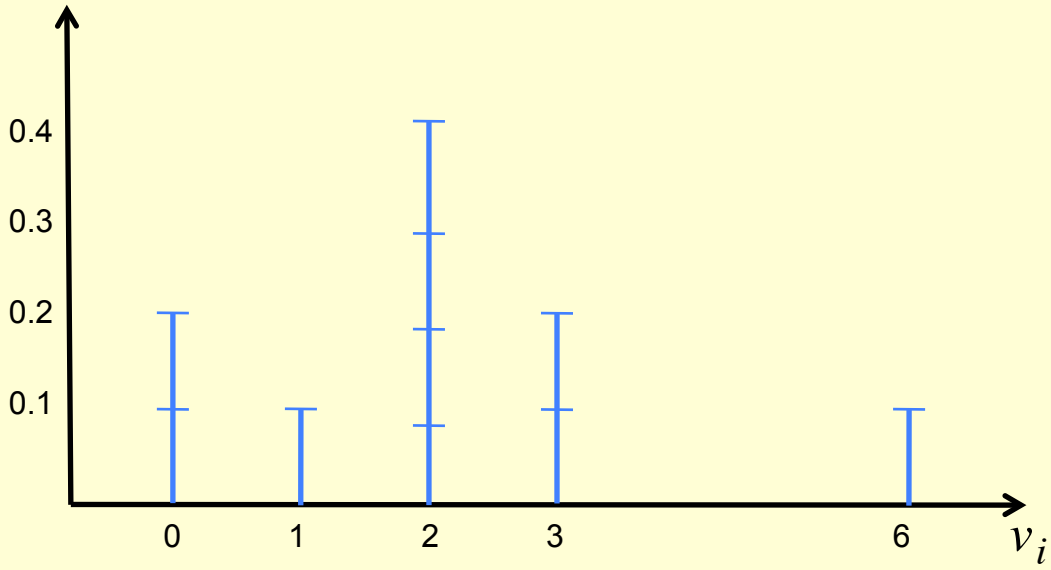
$$\bar{z} = \sum_{i=1}^{5} z_i f_i = 72{,}1\ dmm \rightarrow \bar{x} = 29\ cm + 0.721\ cm = 29.721\ cm$$

Lo stesso discorso si può fare il 7 !

Lo stesso discorso si può fare il 7 ! Non perdiamo tempo a scriverlo: ~~7~~ (*mm*)

v_i	$f_i = \frac{n_i}{n}$
0	0.2
1	0.1
2	0.4
3	0.2
6	0.1

dmm



$$\bar{v} = \sum_{i=1}^{5} v_i f_i = 2.1\ dmm \rightarrow \bar{x} = 29\ cm + 0.021\ cm = 29.721\ cm$$

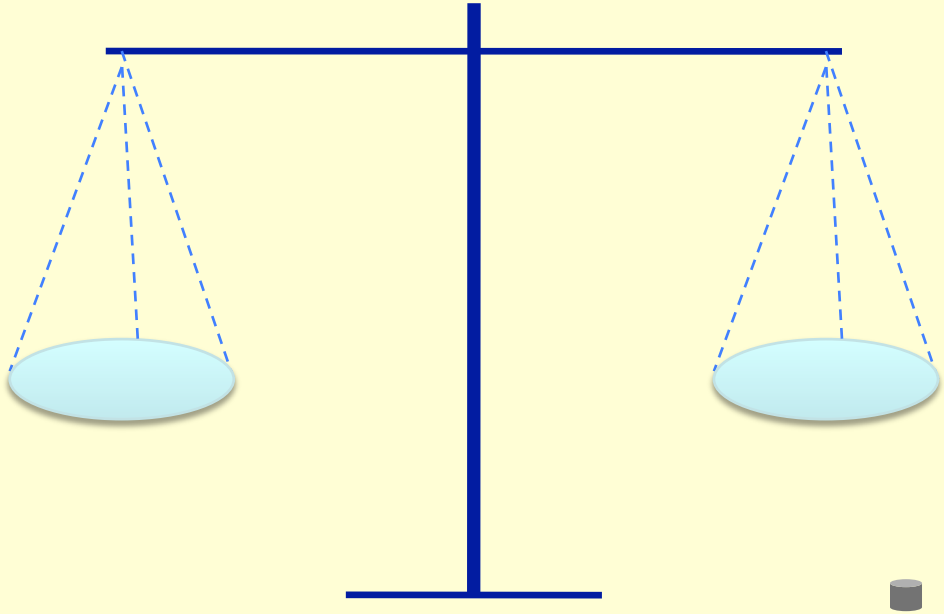
$$(29721\ cmm)$$

fine digressione

abbiate pazienza:
una ulteriore digressione

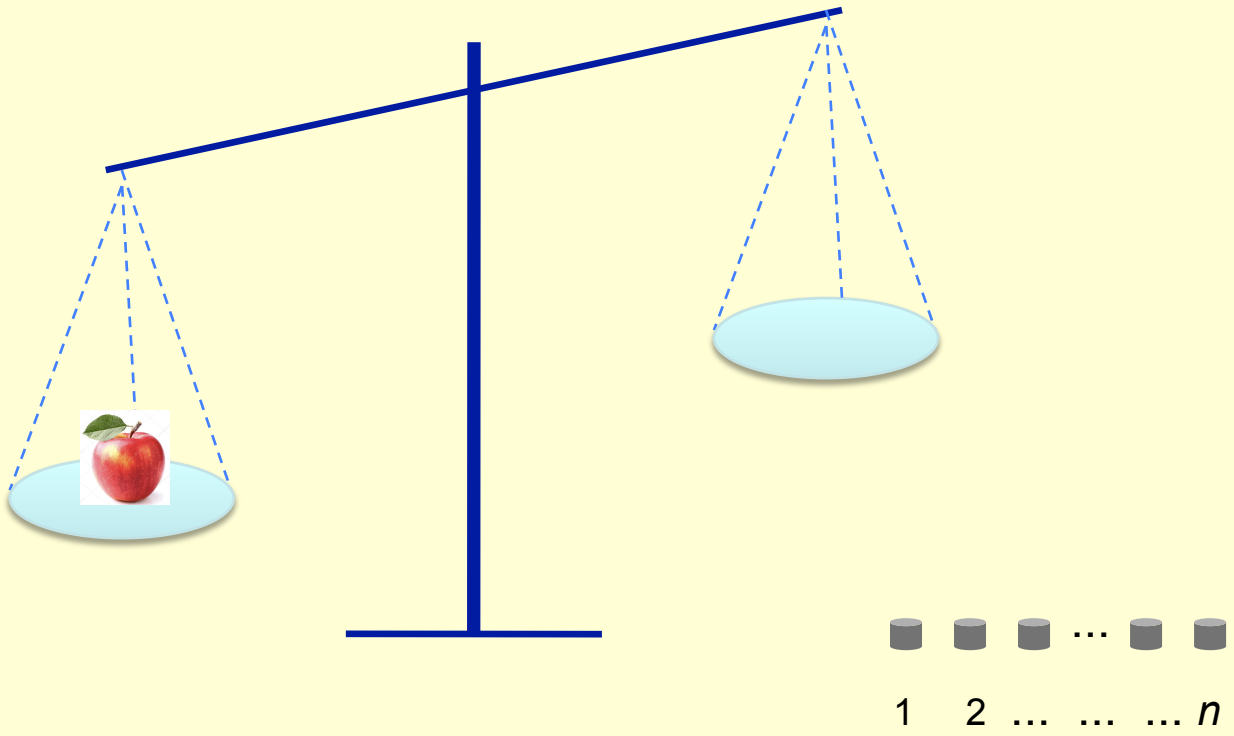
Non sempre si riesce a dare un valore numerico a una misura.

Supponiamo di dover misurare il peso di una mela e di avere a disposizione dei pesetti che supponiamo unitari

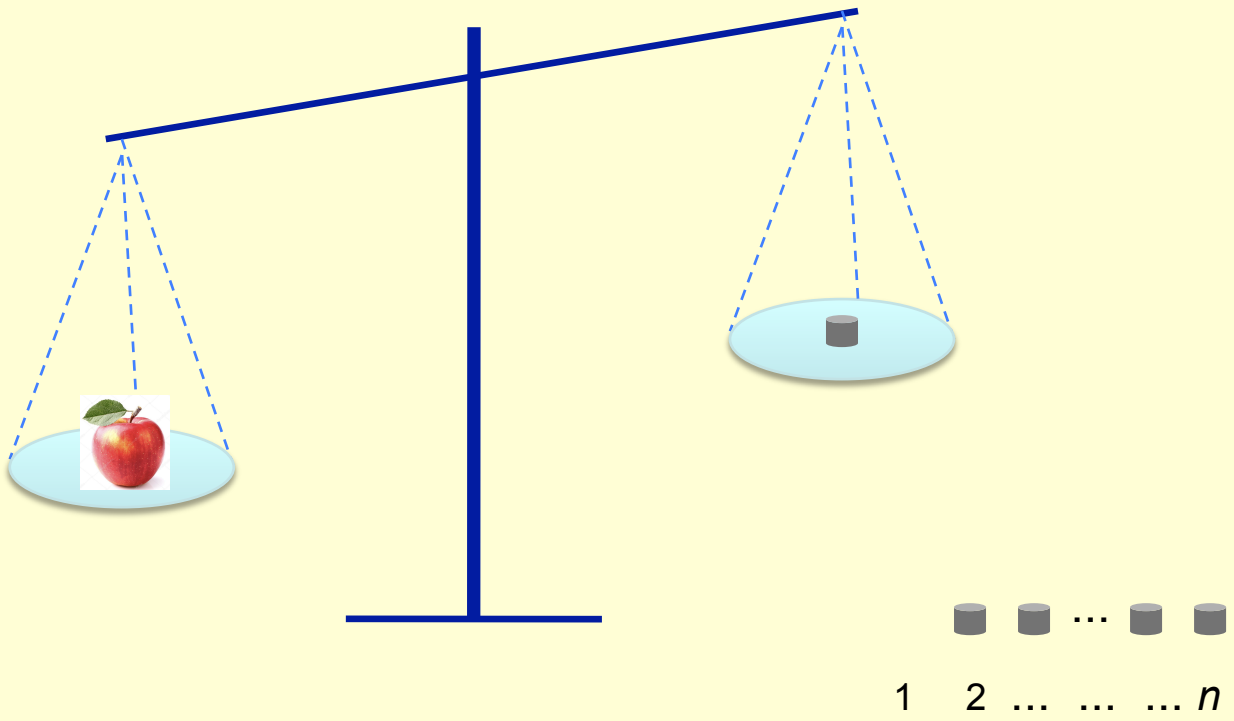


1 2 ... ... *n*

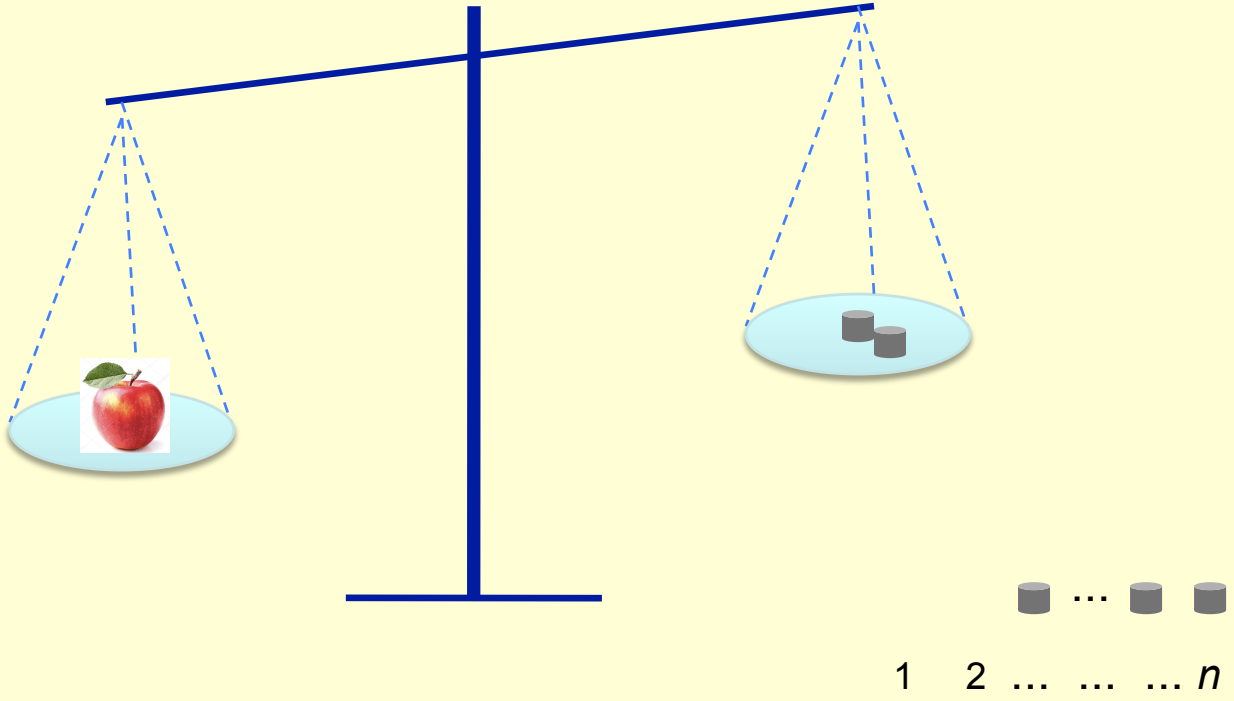
peso di una mela



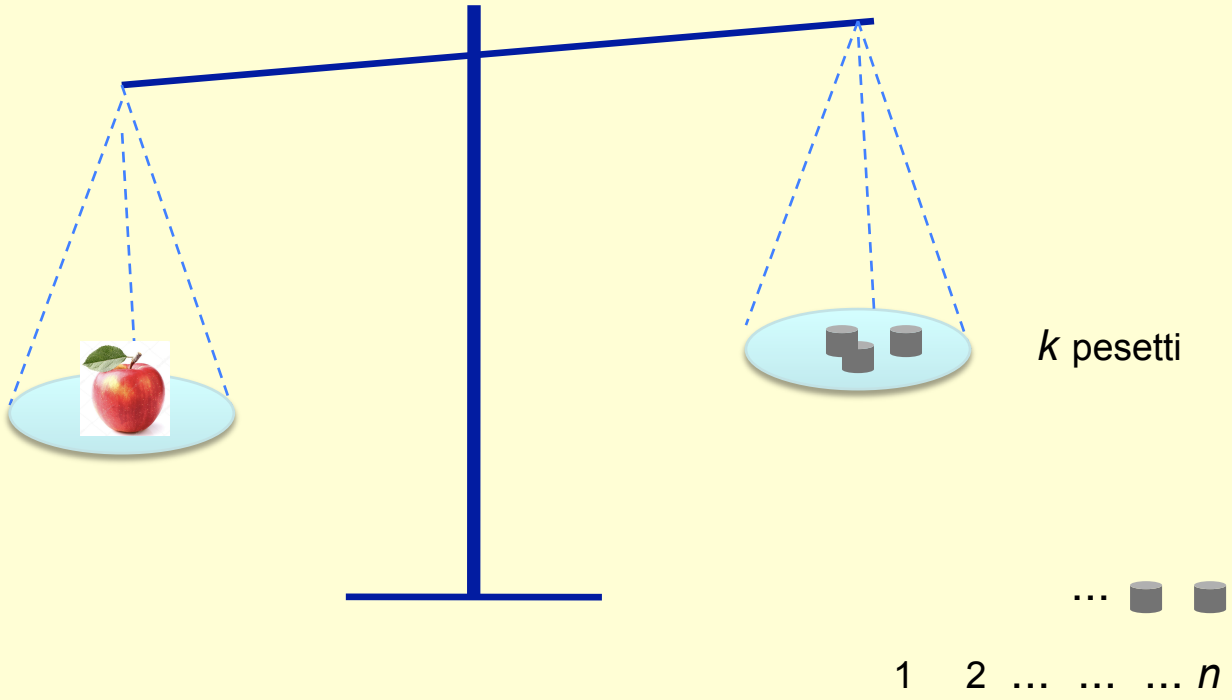
peso di una mela



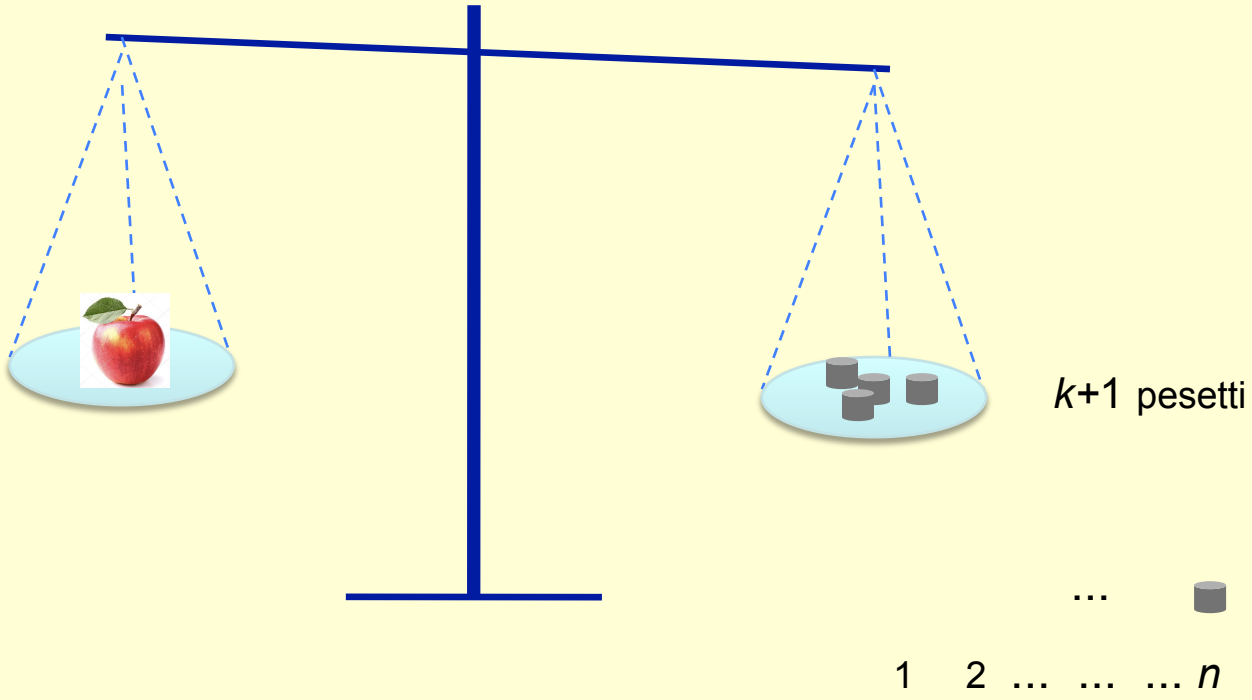
peso di una mela



peso di una mela

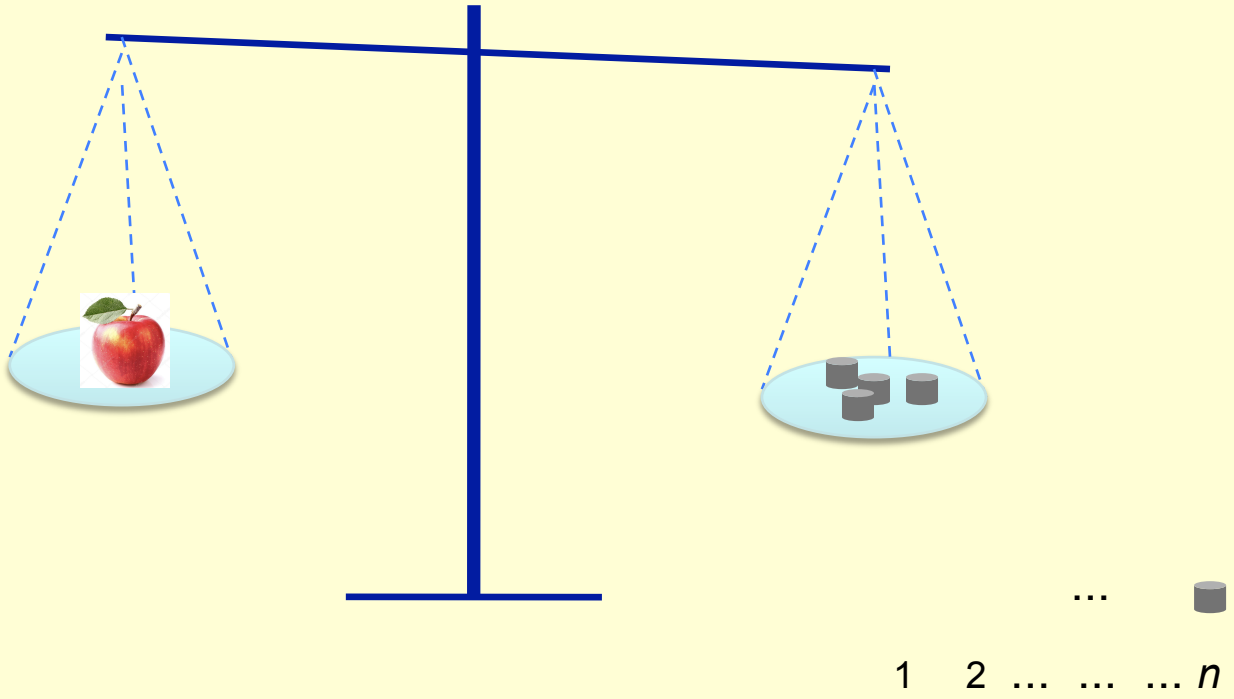


peso di una mela



peso di una mela

Il peso della mela è compreso tra k e k+1 “pesetti”



Anche senza aver definito il peso ($P = mg$), siamo in grado di affermare che

$$P(k) < P(k+1)$$

Questo accade anche per le misure di altre grandezze, ad es.

Temperatura

Il corpo *A* è più **caldo** del corpo *B* : $T(A) > T(B)$

Probabilità

L'evento *A* è più **probabile** dell'evento *B* : $Prob(A) > Prob(B)$

Ciascuno di noi è perfettamente in grado di esprimersi in questo senso

fine digressione

riprendiamo la sintesi di dati raccolti da misure

Supponiamo che i dati rilevati siano numerosi

Es. altezza (in *m*) dei 20 alunni di una classe:

1.54	1.58	1.45	1.60	1.62	1.60	1.55	1.70	1.58	1.61
1.72	1.48	1.43	1.62	1.54	1.60	1.56	1.65	1.47	1.74

E' allora utile (e conveniente) **raggruppare** i dati **in classi**.

Nota

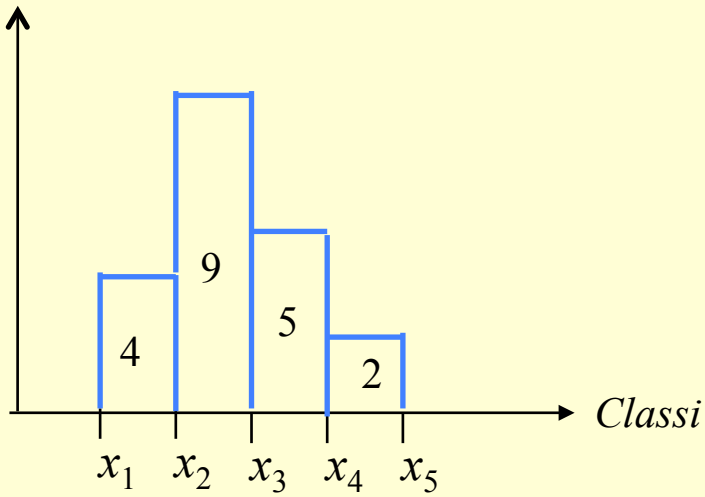
I dati osservati sono sempre discreti, sia che riguardino un **carattere discreto** sia un **carattere continuo.**

Tabella delle frequenze

Classi $(x_i\ ,\ x_{i+1}]$	n_i	f_i
1.40 – 1.50	4	0.20
1.50 – 1.60	9	0.45
1.60 – 1.70	5	0.25
1.70 – 1.80	2	0.10

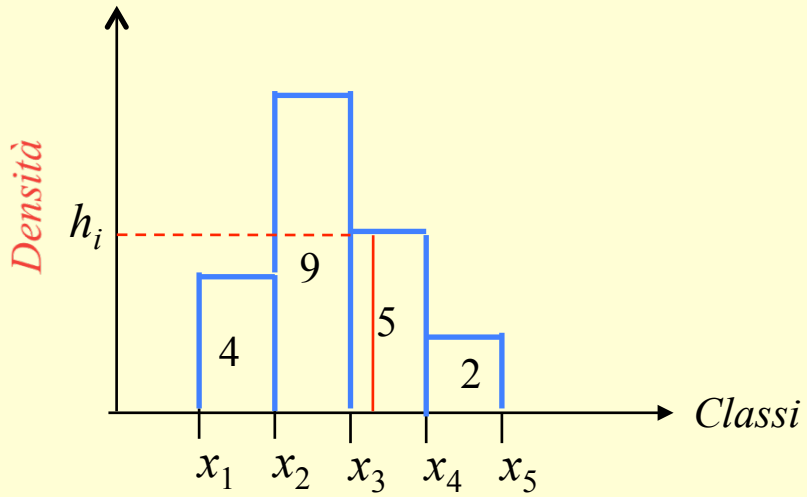
$n=20$

Rappresentazione grafica: **istogramma**



Area rettangolo = frequenza della classe $(n_i$ o $f_i)$

Rappresentazione grafica: **istogramma**

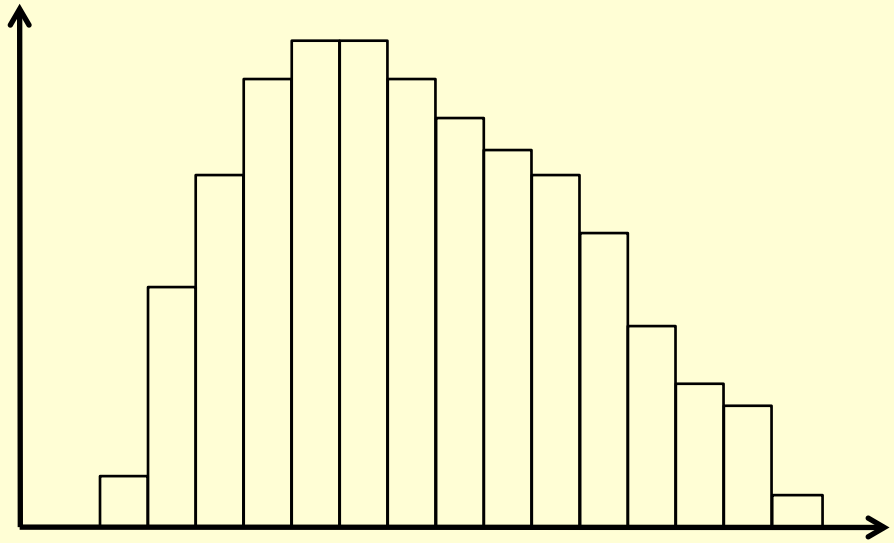


Area rettangolo = frequenza della classe (n_i o f_i)

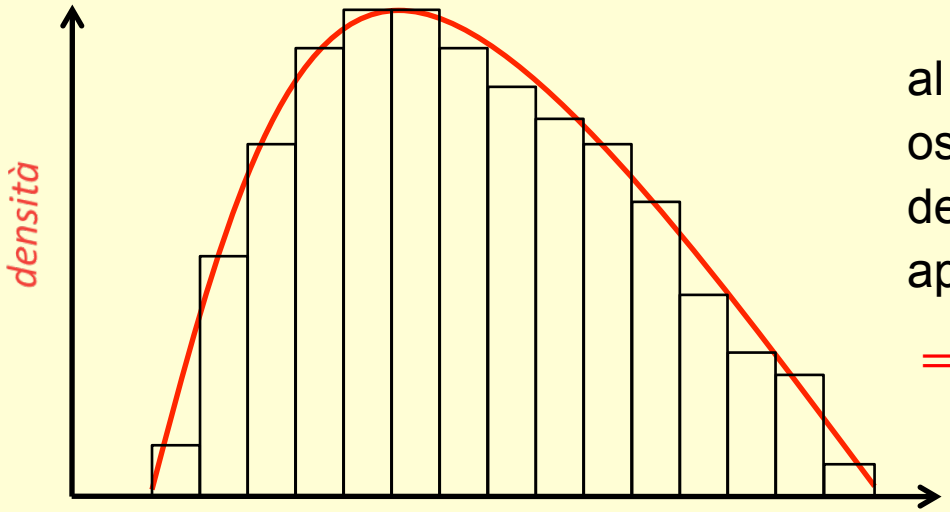
Altezza del rettangolo = densità di <u>frequenza</u>

$$b_i \times h_i = f_i \rightarrow h_i = \frac{f_i}{b_i}$$

Se i dati osservati sono molto numerosi e le classi molto piccole



area rettangoli = f_i
altezza rettangoli = *densità*



al crescere del numero delle osservazioni e, conseguentemente, degli intervalli (classi), l'istogramma è approssimato da una curva liscia

$\Longrightarrow$ *funzione di densità di frequenza*

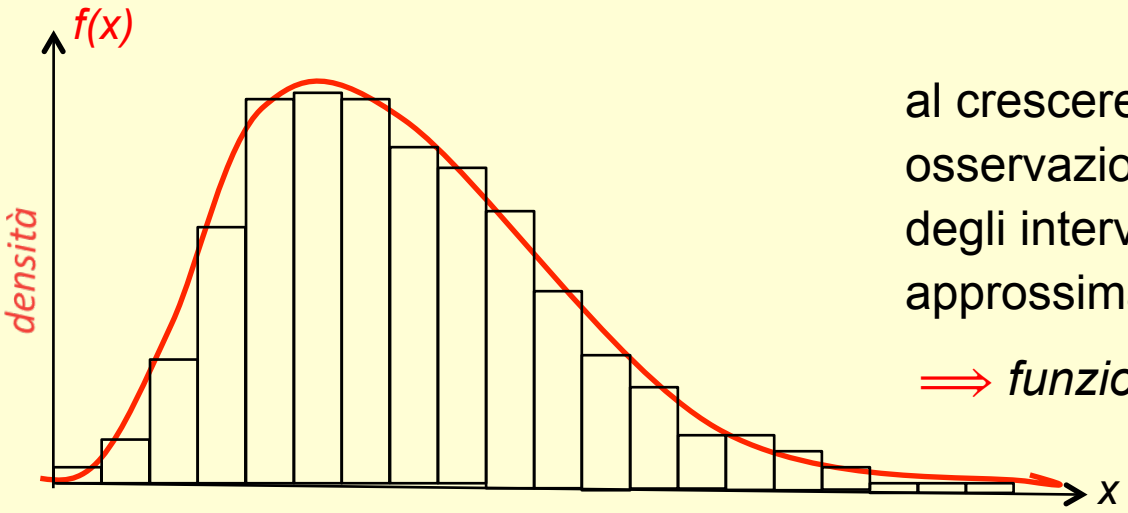
Con la vostra complicità, mi permetto di fare un grande salto di percorso, per affermare come cosa nota che la frequenza relativa f_i , all'aumentare nel numero delle osservazioni, è una approssimazione della probabilità.

Di conseguenza la *funzione di densità delle frequenze* può essere riguardata come *funzione di densità della probabilità*.

Nel caso di un carattere continuo

Se “infittiamo” gli intervalli

Rappresentazione grafica: *funzione di densità della probabilità*

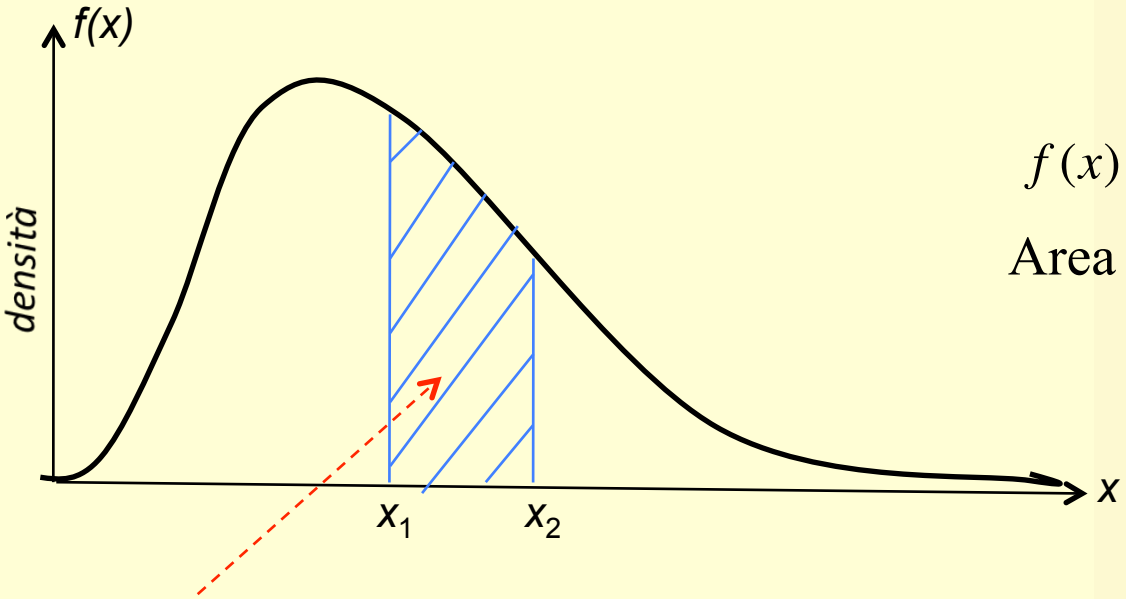


al crescere del numero delle osservazioni e, conseguentemente, degli intervalli (classi), l’istogramma è approssimato da una curva liscia *f(x)*

$\Longrightarrow$ *funzione di densità di probabilità*

Distribuzione di probabilità di una generica v.c. continua

Proprietà della funzione di densità *f(x)*



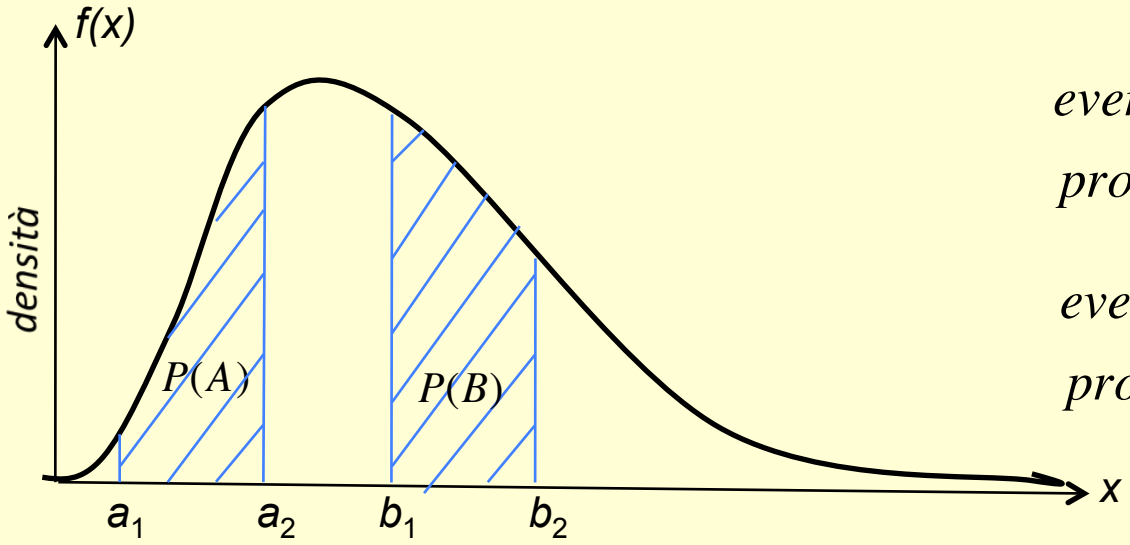
$f(x) \geq 0$

Area totale sotto la curva = 1

Prob. che x sia compresa tra x_1 e x_2 = $P(x_1 < x < x_2)$ = *area tratteggiata*

Formalmente: $P(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)dx$

Proprietà della funzione di densità *f(x)*



$evento\ \ A : (a_1 \le x \le a_2)$

$probabilità\ \text{di}\ A : P(A)$

$evento\ \ B : (b_1 \le x \le b_2)$

$probabilità\ \text{di}\ B : P(B)$

Dal confronto tra le aree: $P(A) < P(B)$

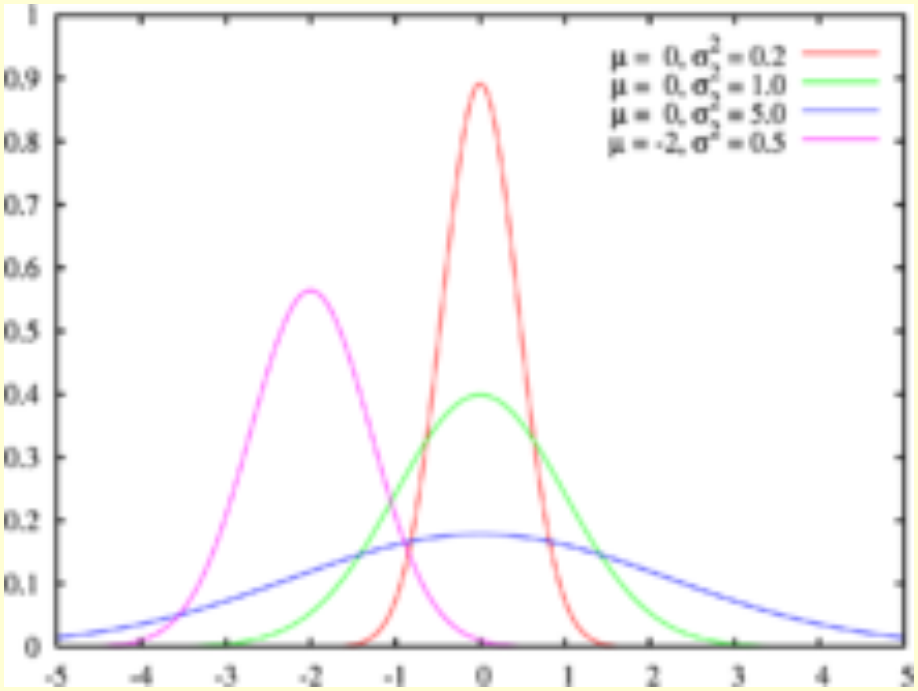
In una rilevazione futura il valore di x “cadrà” in B più probabilmente che in A

abbiamo ora gli strumenti per

misurare l'incertezza delle misure

La curva degli errori di Gauss (*Gaussiana* o *Normale*)

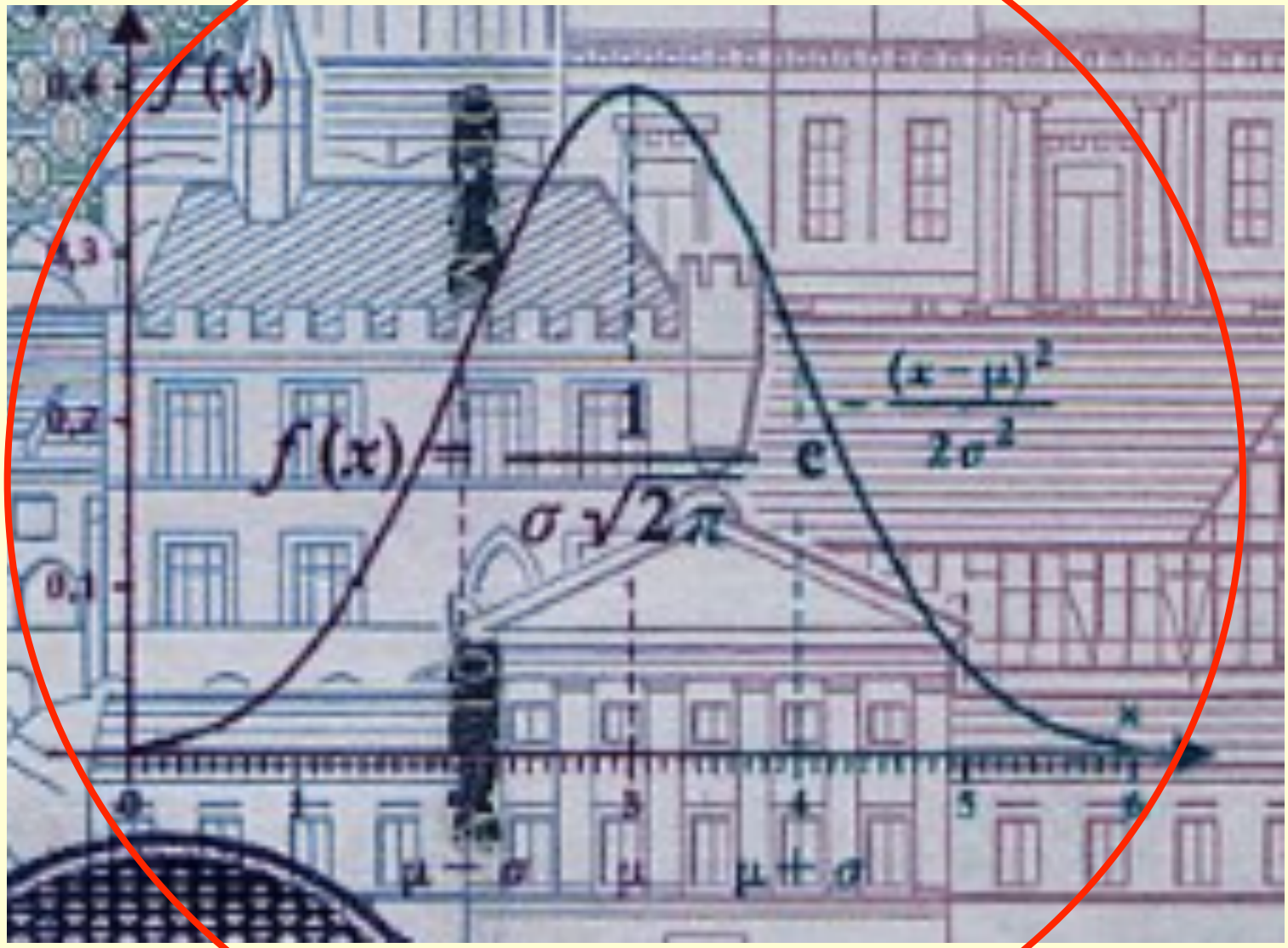
Carl Friedrich Gauss (1777-1855)



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty \qquad -\infty < \mu < \infty \qquad \sigma > 0$$


GN4480100S8
DEUTSCHE BUNDESBANK
Banknote
10
1777–1855 Carl Friedr. Gauß
ZEHN DEUTSCHE MARK
GN4480100S8
Deutsche Bundesbank
Frankfurt am Main
1. September 1999


GN4480100S8
DEUTSCHE BUNDESBANK
Banknote
10
1777–1855 Carl Friedr. Gauß
ZEHN DEUTSCHE MARK
GN4480100S8
Deutsche Bundesbank
Frankfurt am Main
1. September 1999


f(x)
0,4
0,3
0,2
0,1
f(x) = 1/(σ√2π) e^(−(x−μ)²/2σ²)
x
μ − σ
μ
μ + σ

La curva degli errori accidentali
(Curva di Gauss; Curva di Gauss-Laplace)
Distribuzione di probabilità *Normale (*o *Gaussiana)*

- La prima formulazione è attribuibile a

  Abraham de Moivre (1667 Francia – 1754 Inghilterra) che la costruì nel 1733 ma i suoi scritti andarono persi sino al loro ritrovamento nel 1924 da parte di Karl Pearson (1857 – 1936 Londra) che gliene restituì il merito e la denominò estensivamente con il termine "*normale*" già coniato da altri (Charles Sanders Peirce, Wilhelm Lexis, Francis Galton intorno al 1875)

- Nel 1783 Pierre Simon de Laplace (1749 – 1827 Francia) la utilizzò per descrivere la distribuzione degli errori accidentali di misura.

- Nel 1809 Carl Friedrich Gauss (1777 – 1855 Germania) la utilizzò per lo studio di dati astronomici e ne approfondì e divulgò le proprietà

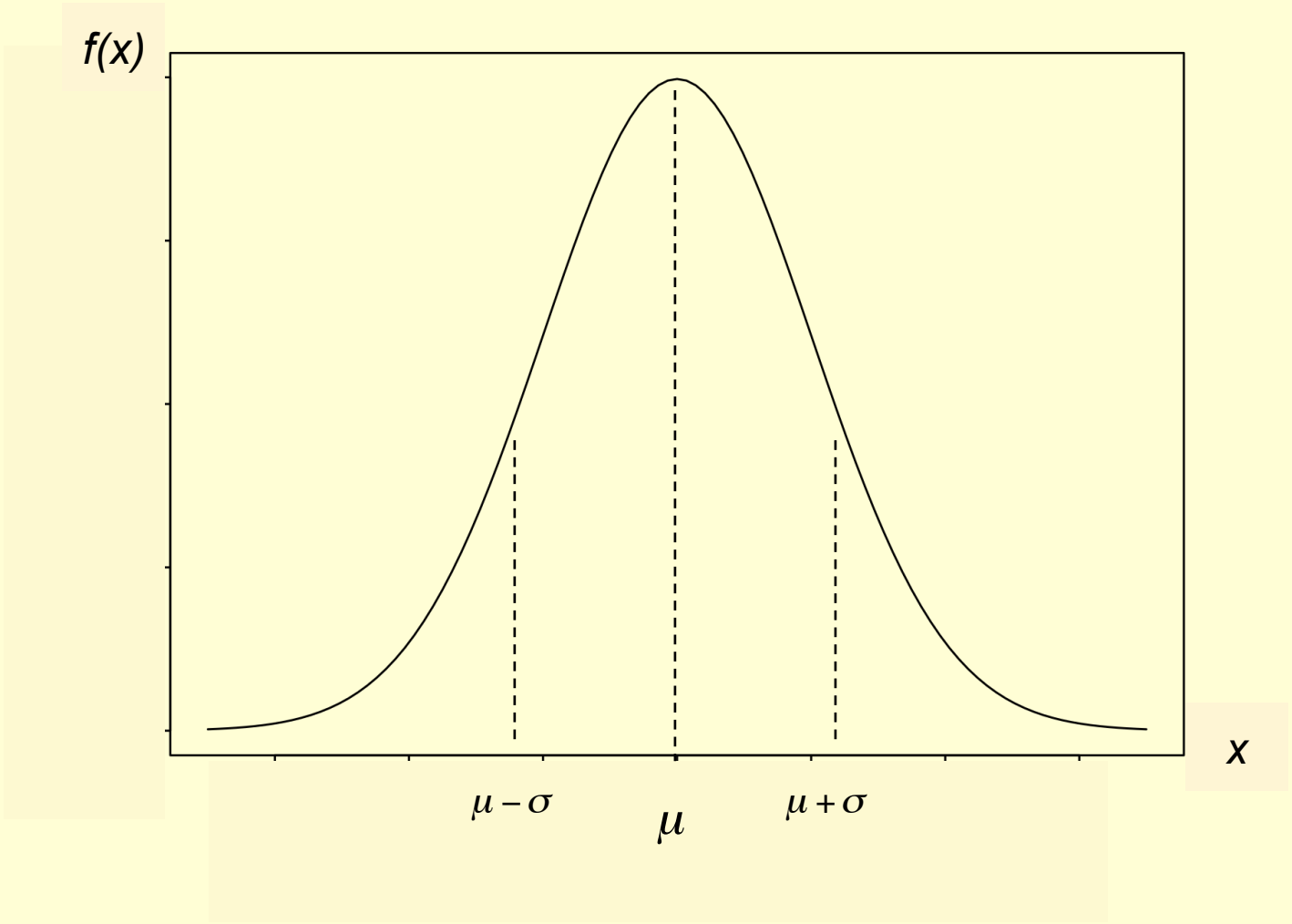
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty \;;\; -\infty < \mu < \infty \;;\; \sigma > 0$$

La funzione $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Tralasciandone la genesi e la natura probabilistica e riguardando la *f(x)* come una funzione reale di variabile reale, possiamo studiarla e costruirne il grafico

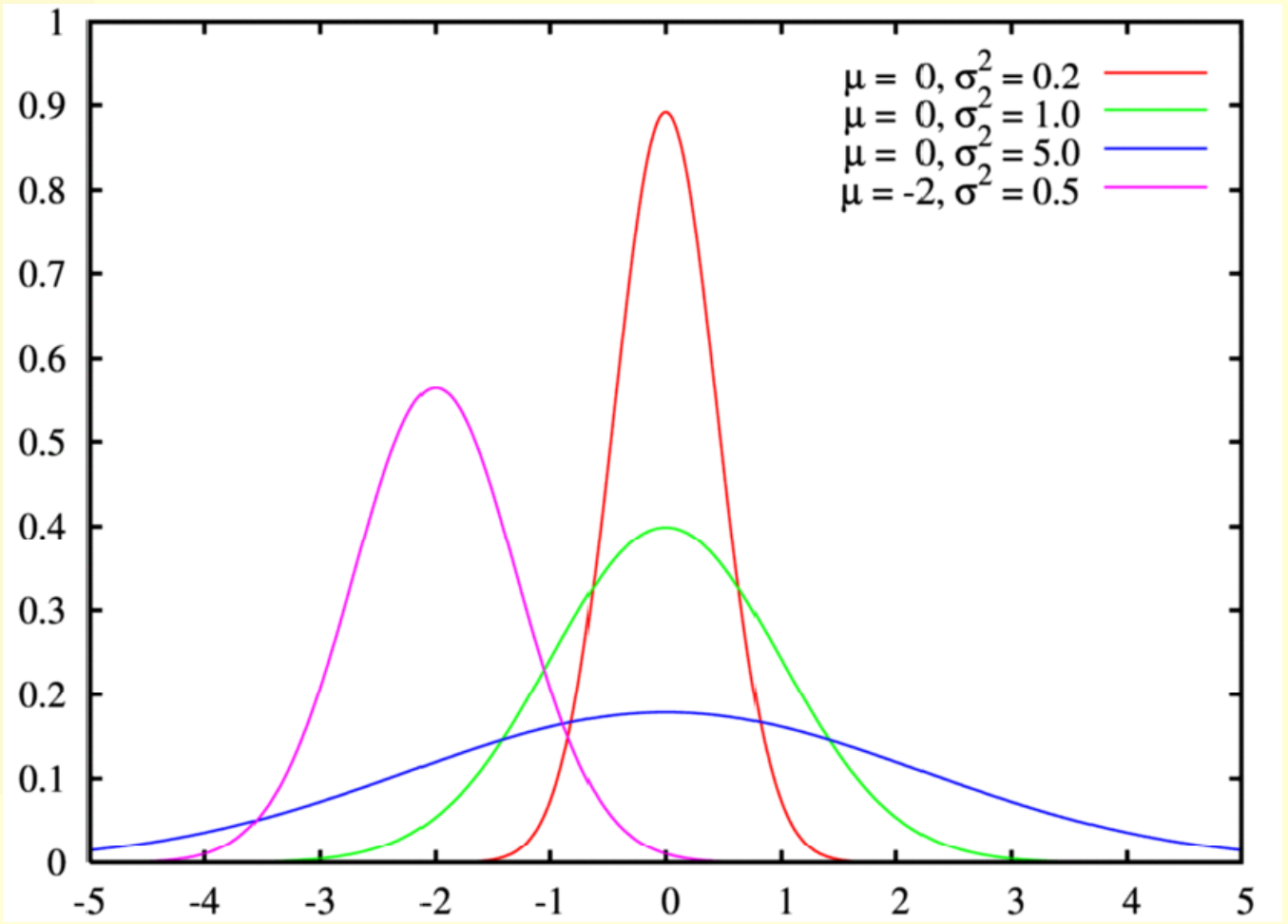
- *campo di definizione* : $-\infty < x < \infty$
- *positiva* : $f(x) \geq 0 \quad , \ \forall\ x$
- *simmetrica rispetto all'asse* $x = \mu$: $f[(x-\mu)] = f[-(x-\mu)]$
- *asintoto orizzontale* $y = 0$: $\lim_{x \to \pm\infty} f(x) = 0$
- *punto di massimo in* $x = \mu$: $\max_x f(x) = f(\mu)$
- *flessi ascendente e discendente, rispettivamente, in* $x = \mu \mp \sigma$

La funzione $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



La funzione $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

dipendenza dai parametri μ e σ



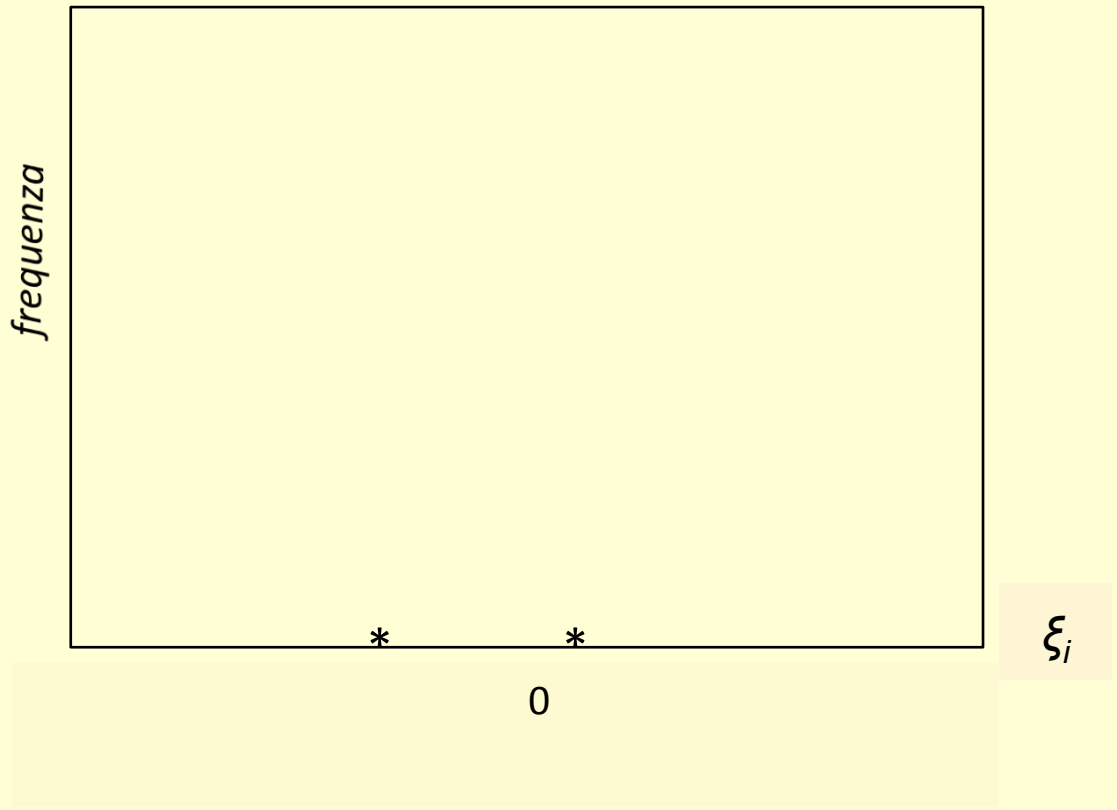
La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

n misure *ripetute* su una grandezza di valore incognito *μ (lunghezza, peso, tempo, ecc.*), denotando i valori osservati con $x_i = \mu + \xi_i$ (i =1, 2, ..., n)

e con ξ_i *errori accidentali* di misura

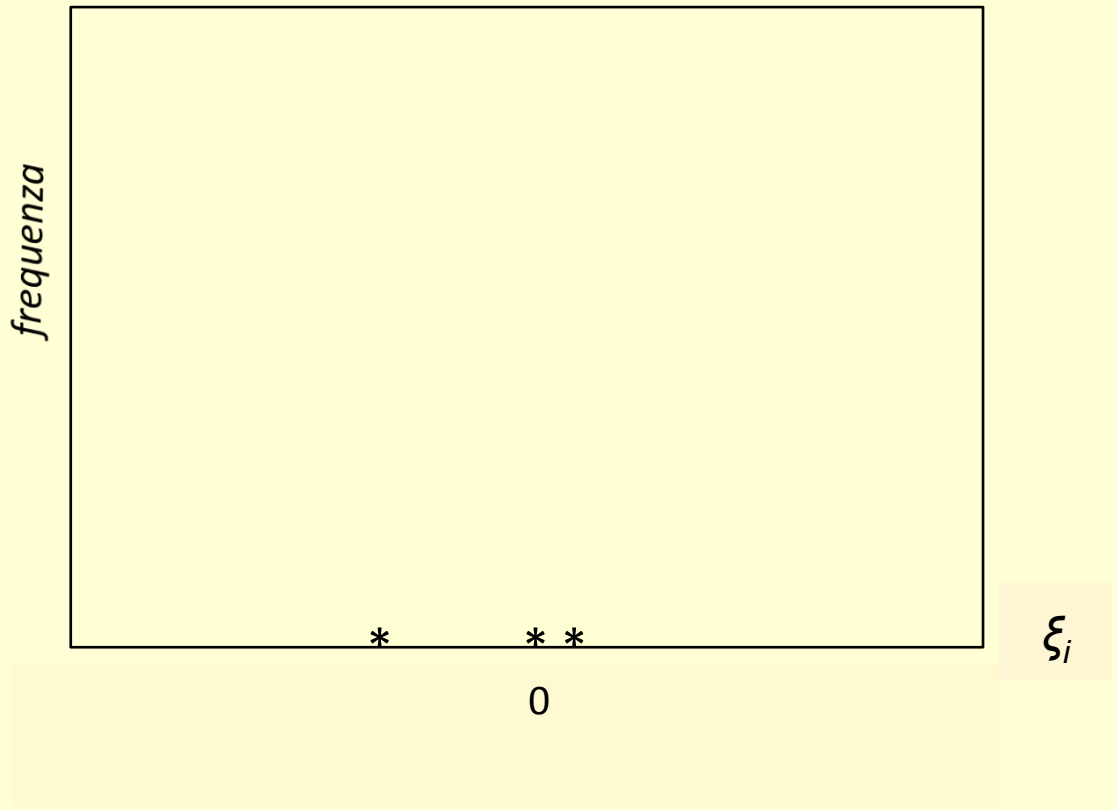
La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

n misure *ripetute* su una grandezza di valore incognito *μ (lunghezza, peso, tempo, ecc.*), denotando i valori osservati con $x_i = \mu + \xi_i$ (i =1, 2, ..., n)
e con ξ_i *errori accidentali* di misura



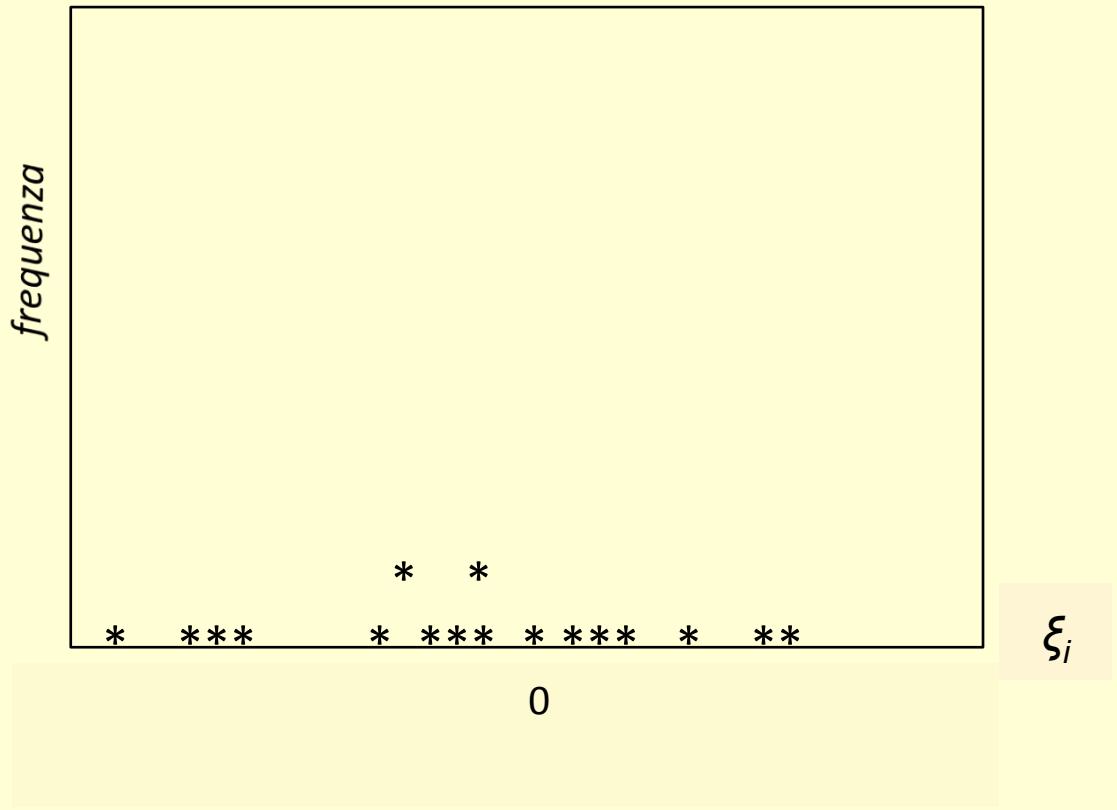
La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

n misure *ripetute* su una grandezza di valore incognito μ *(lunghezza, peso, tempo, ecc.*), denotando i valori osservati con $x_i = \mu + \xi_i$ (i =1, 2, ..., n)
e con ξ_i *errori accidentali* di misura



La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

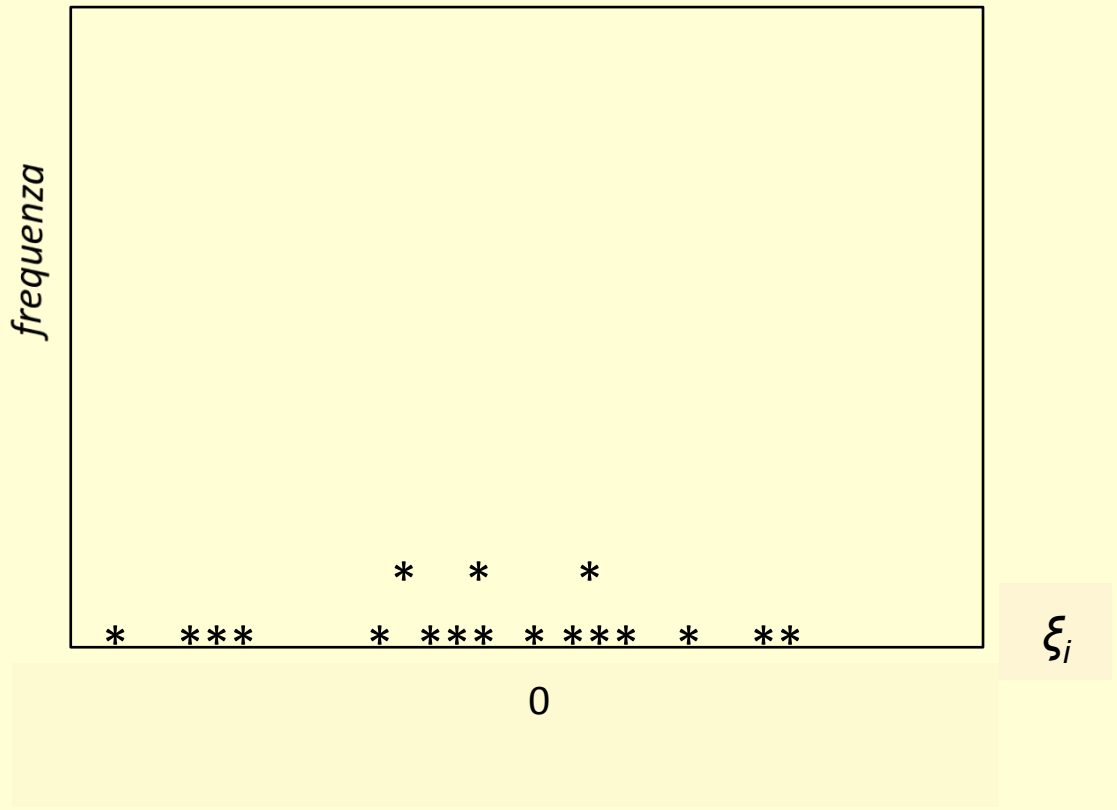
n misure *ripetute* su una grandezza di valore incognito *μ (lunghezza, peso, tempo, ecc.*), denotando i valori osservati con $x_i = \mu + \xi_i$ (i =1, 2, ..., n)
e con ξ_i *errori accidentali* di misura



La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

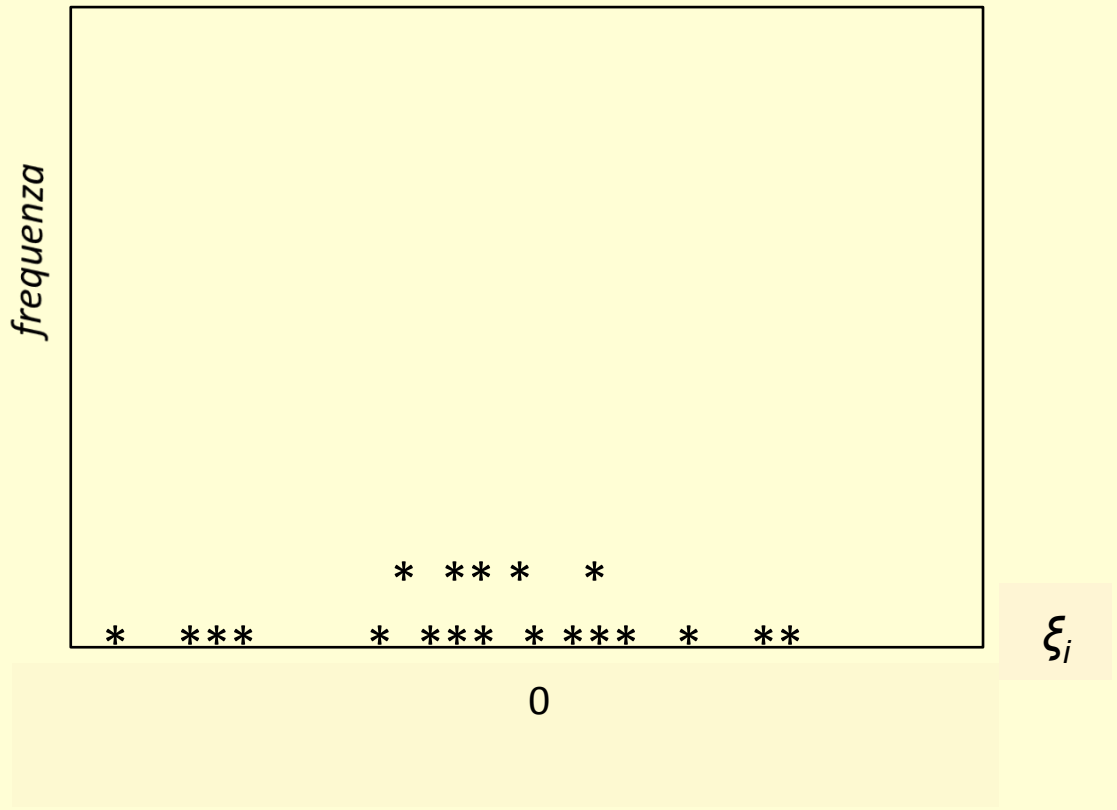
n misure *ripetute* su una grandezza di valore incognito *μ (lunghezza, peso, tempo, ecc.*), denotando i valori osservati con $x_i = \mu + \xi_i$ (i =1, 2, ..., n)

e con ξ_i *errori accidentali* di misura



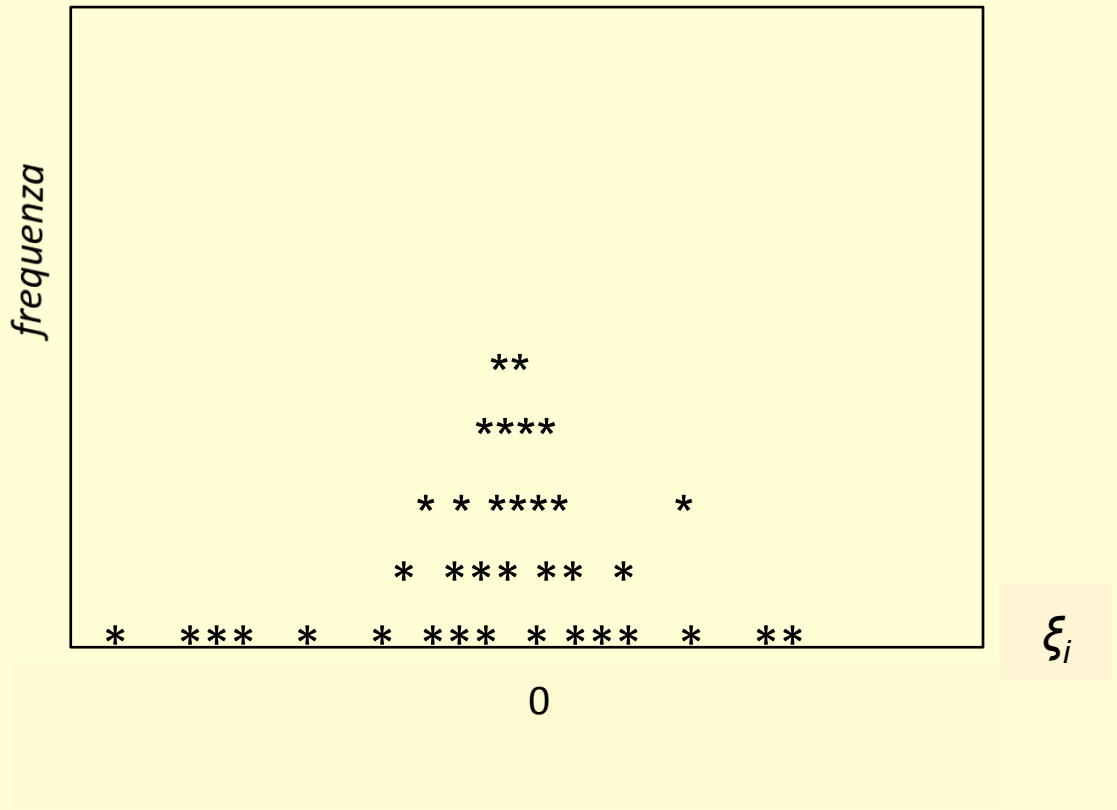
La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

n misure *ripetute* su una grandezza di valore incognito *μ (lunghezza, peso, tempo, ecc.*), denotando i valori osservati con $x_i = \mu + \xi_i$ (i =1, 2, ..., n)
e con ξ_i *errori accidentali* di misura



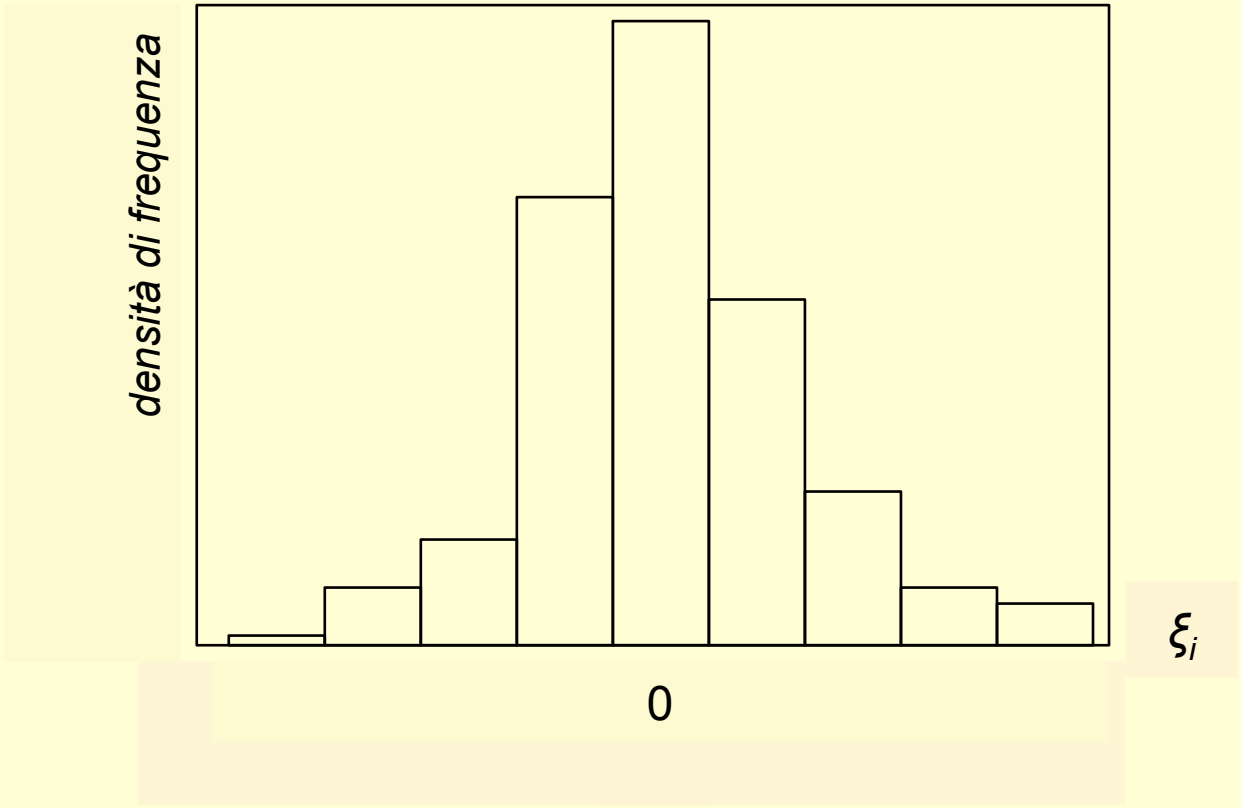
La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

n misure *ripetute* su una grandezza di valore incognito *μ (lunghezza, peso, tempo, ecc.*), denotando i valori osservati con $x_i = \mu + \xi_i$ (i =1, 2, ..., n)
e con ξ_i *errori accidentali* di misura



La funzione di densità di probabilità $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

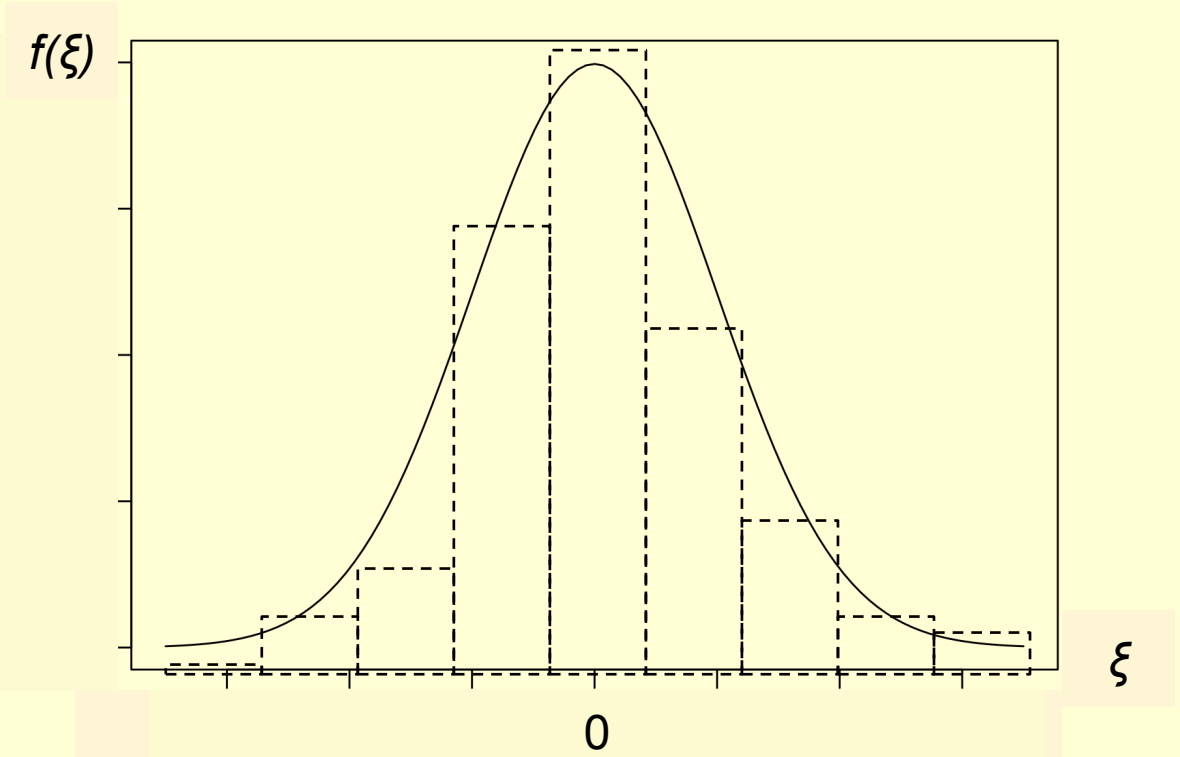
all'aumentare del numero *n* delle misure è opportuno raggruppare i dati in classi e costruire l'istogramma delle frequenze



Infittendo gli intervalli l'istogramma tende ad "assumere la forma" di una *curva gaussiana* con media (valore atteso) 0 e deviazione standard σ.

Nota: gli errori ξ sono *puramente accidentali*: se fosse $\mu \neq 0$ ci sarebbe una componente d'errore sistematica. σ rappresenta inversamente la precisione della misura.

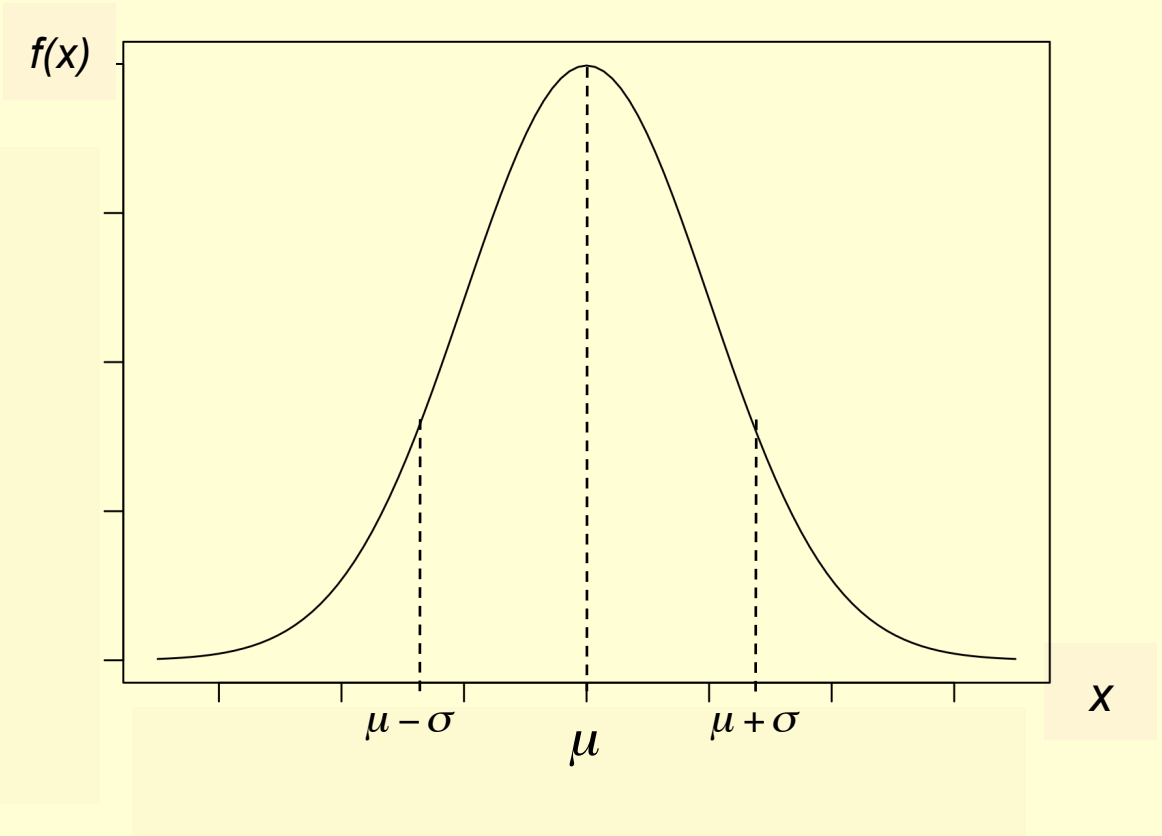
$$f(\xi) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\xi^2}{2\sigma^2}}$$



Dal modello probabilistico degli errori *f(ξ)* si ottiene di conseguenza la distribuzione di probabilità dei valori di misura $x_i = \mu + \xi_i \rightarrow (\xi = x - \mu)$

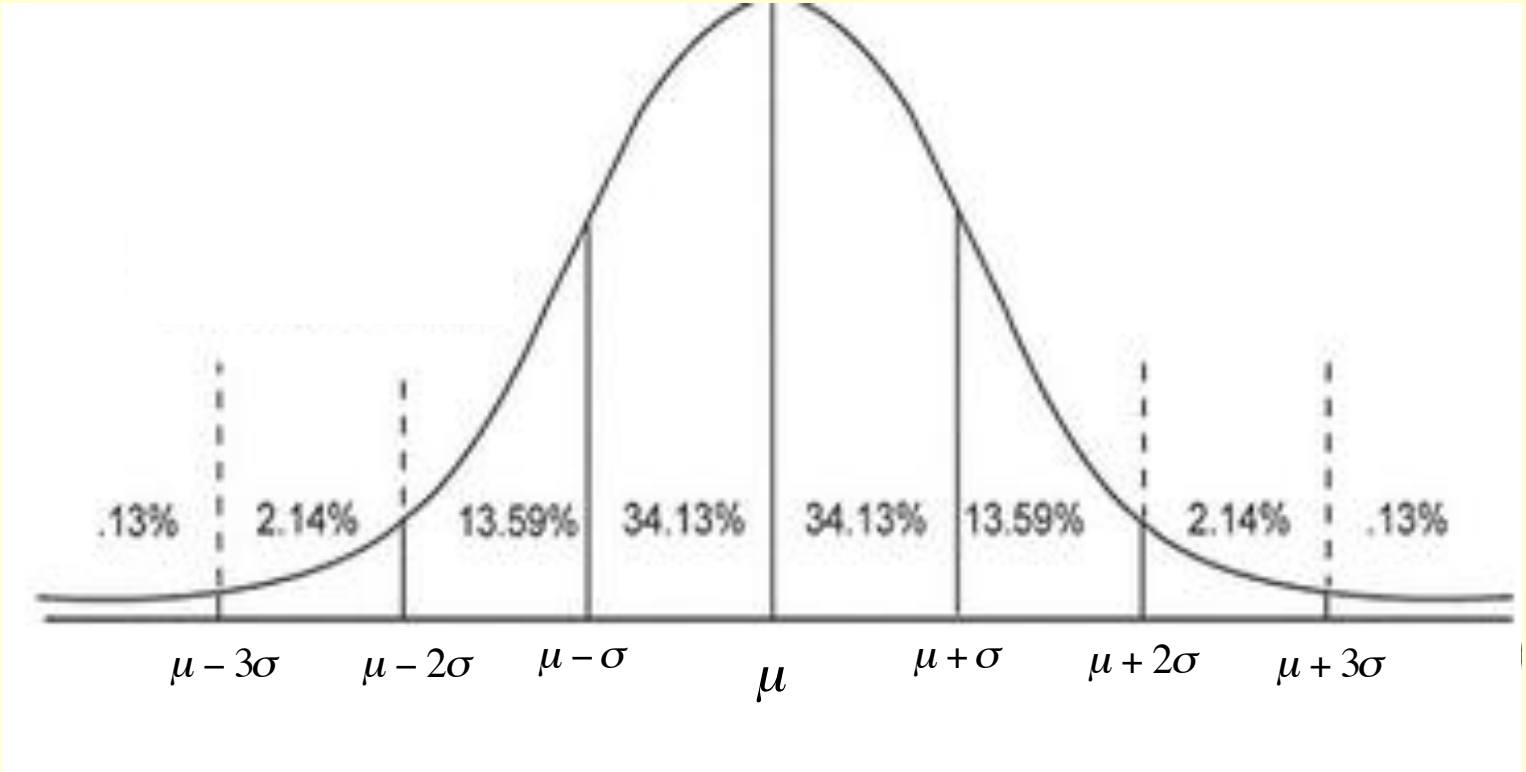
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

con media μ e *sd*= σ



I piccoli errori sono i più probabili

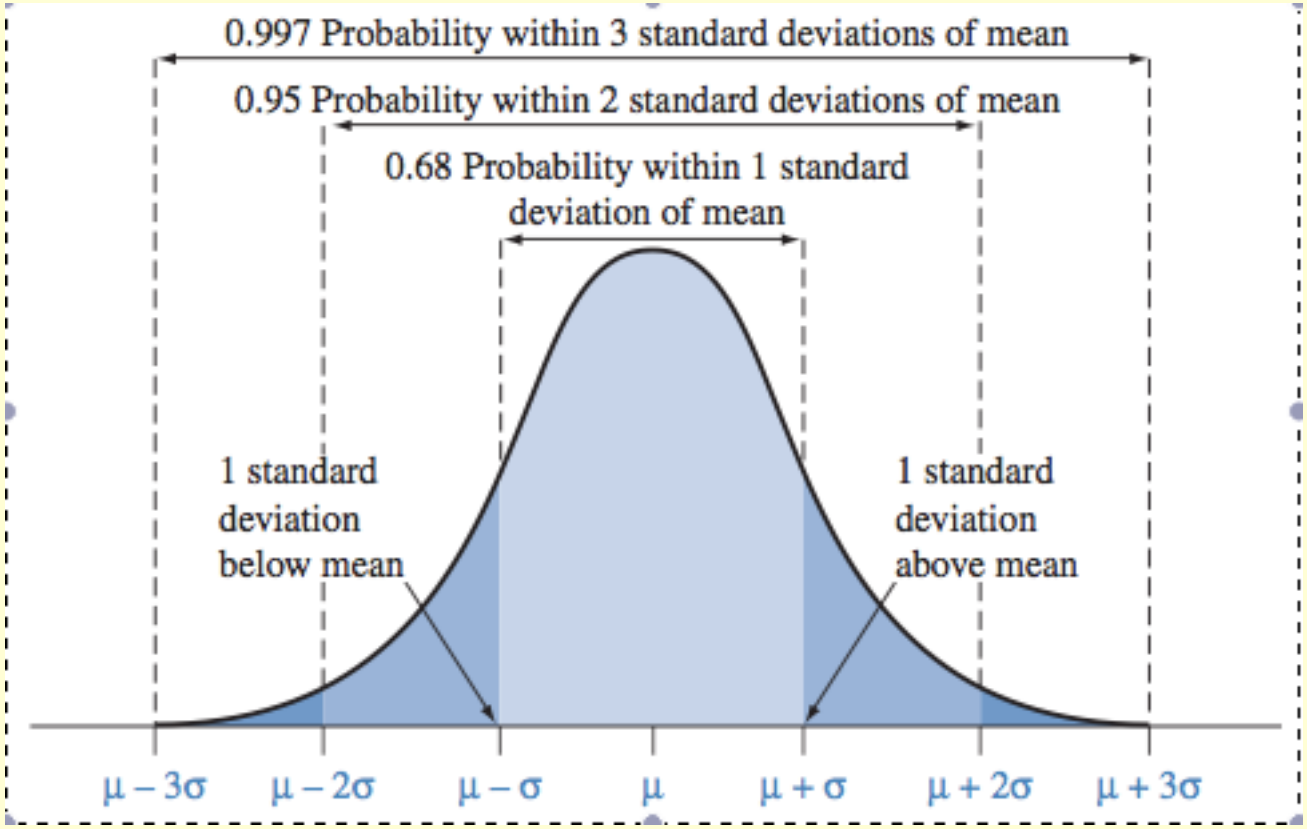
Aree sotto la curva (= probabilità)



Osservazione. I valori di *x* compresi in un intorno della media di 3*σ raccolgono* il 99.72% della probabilità: questo spiega perché, anche se per definizione *x* varia tra -∞ e +∞, il modello può rappresentare grandezze positive come lunghezze, pesi, tempi, ecc.

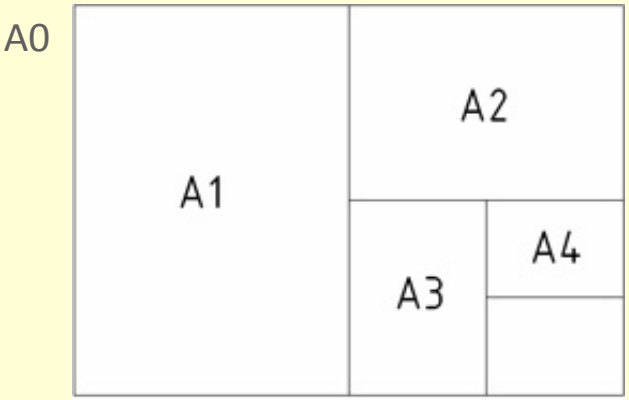
Aree sotto la curva (= probabilità)

(*in sintesi*)



Una osservazione che cade a una distanza > di 3σ dalla media può essere considerata come un "*valore anomalo*" rispetto al modello adottato

Interpretare un’area come misura della probabilità di un evento è intuitivo:
si immagini di lanciare una monetina o un sassetto su un foglio di carta A3
(è soltanto un esempio!)



- A0 è l'***evento certo*** giacché si assume che la moneta non possa cadere al di fuori;
- i sottoinsieme tracciati costituiscono dunque una partizione dell'*evento certo*;
- la prob. che la moneta cada su A1 è pari alla metà dell'area totale che assumiamo = 1 (*evento certo*!);
- la prob. che cada su A4 è allora pari a 1/16

In sintesi

- Rappresentazione di eventi mediante insiemi (*diagrammi di Venn*)
- Probabilità = area dell'insieme che rappresenta l'evento, rispetto all'area totale (pari a 1 per l'evento certo)

al cinema

(film di una sola scena, 2 azioni + un epilogo)

Nel saloon

Ambientazione

un saloon del far west con diversi avventori.

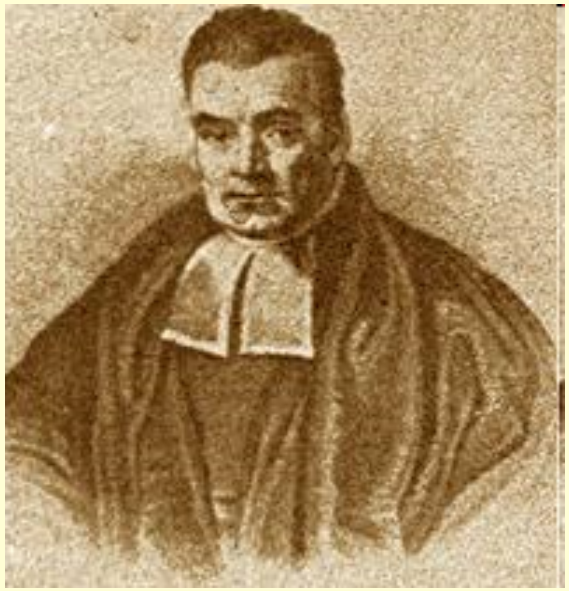
Azione 1

S’aprono le porte del saloon, entra il cow boy *Piero*. Ha una faccia da duro.

Piero estrae dalla fondina un mazzo di carte francesi e dice “tu smazzi, se la carta è rossa vinci, se è nera vinco io. Chi gioca?”

Fine primo tempo

Intervallo pubblicitario



Reverendo Thomas Bayes (1702-1761)

Teorema di Bayes

Che insegna ad aggiornare la valutazione della probabilità di un evento alla luce di nuove informazioni, oltre quelle iniziali.

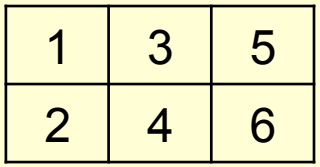
Prima del

Secondo tempo

è necessario fare una premessa

Premessa

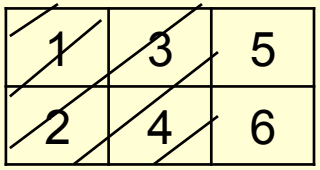
Lancio di un dado



Ω Evento certo

$P(\Omega) = 1$

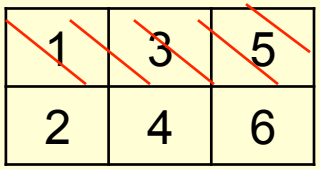
Evento $E : \{x < 5\} = \{1,2,3,4\}$



E

$P(E) = \frac{4}{6} = \frac{2}{3}$

Evento $A : \{dispari\} = \{1,3,5\}$



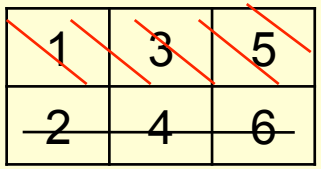
A

$P(A) = \frac{3}{6} = \frac{1}{2}$

lancio $\longrightarrow$ esito *dispari* (si verifica A), qual è la probabilità di E ?

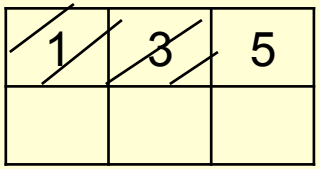
Qual è la prob. che "essendo uscito un *dispari*, esso sia $x < 5$" ? $P(E \mid A)$?

$P(E \mid A)$?



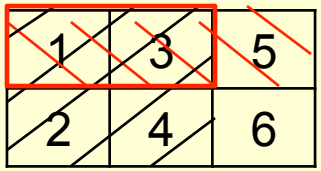
A

$P(A) = \frac{3}{6}$



$E \mid A$

$P(E \mid A) = \frac{2}{3}$



$A \cap E$

$P(A \cap E) = \frac{2}{6}$

In generale $\boxed{P(E \mid A) = \frac{P(A \cap E)}{P(A)}} = \frac{2/6}{3/6} = \frac{2}{3}$ $\boxed{P(A \cap E) = P(A) \cdot P(E \mid A)} = \frac{3}{6} \cdot \frac{2}{3} = \frac{2}{6}$

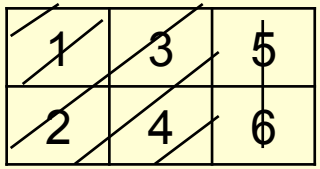
L'idea (semplice ma geniale) di T. Bayes

"*se invece so che si è verificato* $E : \{x < 5\}$ *qual è la probabilità di A*?"

Qual è la prob. che, "essendo uscito un $x < 5$, esso sia *dispari*" ? $P(A \mid E)$?

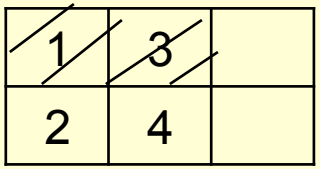
In sintesi: come l'informazione sull'essersi verificato A modifica la valutazione della prob. di E, anche il verificarsi di E modifica la valutazione della prob. di A

$$P(A \mid E)?$$



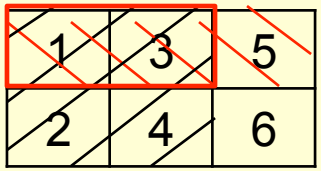
E

$$P(E) = \frac{4}{6}$$



$A \mid E$

$$P(A \mid E) = \frac{2}{4} = \frac{1}{2}$$



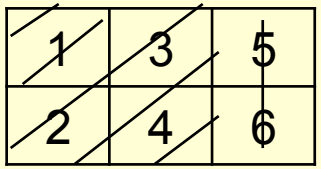
$A \cap E$

$$P(E \cap A) = \frac{2}{6}$$

$$\boxed{P(A \mid E) = \frac{P(A \cap E)}{P(E)}} = \frac{2/6}{4/6} = \frac{1}{2}$$

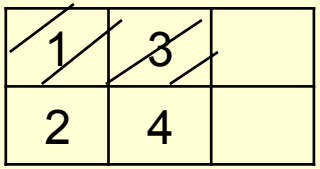
$$\boxed{P(E \cap A) = P(E) \cdot P(A \mid E)} = \frac{4}{6} \cdot \frac{2}{4} = \frac{2}{6}$$

$$P(A \mid E)?$$



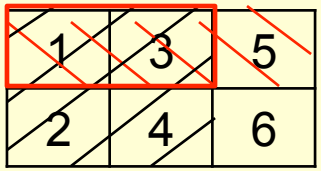
E

$$P(E) = \frac{4}{6}$$



$A \mid E$

$$P(A \mid E) = \frac{2}{4} = \frac{1}{2}$$



$A \cap E$

$$P(E \cap A) = \frac{2}{6}$$

$$\boxed{P(A \mid E) = \frac{P(A \cap E)}{P(E)}} = \frac{2/6}{4/6} = \frac{1}{2} \qquad \boxed{P(E \cap A) = P(E) \cdot P(A \mid E)} = \frac{4}{6} \cdot \frac{2}{4} = \frac{2}{6}$$

Teorema di Bayes

$$\left.\begin{aligned} P(A \cap E) &= P(A) \cdot P(E \mid A) \\ P(E \cap A) &= P(E) \cdot P(A \mid E) \end{aligned}\right\} \quad P(E) \cdot P(A \mid E) = P(A) \cdot P(E \mid A)$$

$$P(A \mid E) = \frac{P(A) \cdot P(E \mid A)}{P(E)}$$

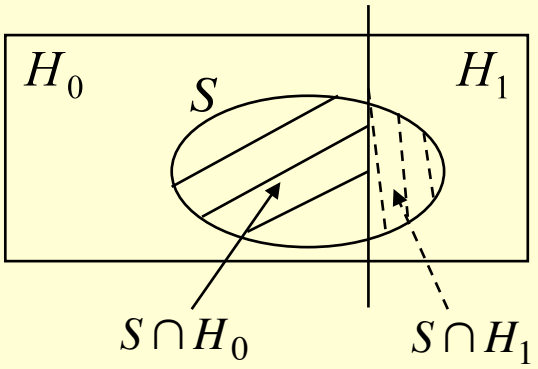
Per capirne meglio l'importanza

(teorema della probabilità delle cause)

All'esame:

lo studente può essere

- *preparato* H_0
- *non-preparato* $H_1 \ (= \overline{H}_0)$



Evento: $S =$ *lo studente supera l'esame*

$$S = (S \cap H_0) \cup (S \cap H_1)$$

$$P(S) = P(S \cap H_0) + P(S \cap H_1) =$$

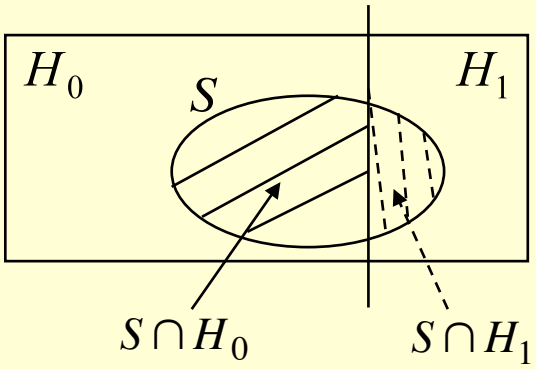
$$= P(H_0) \cdot P(S \mid H_0) + P(H_1) \cdot P(S \mid H_1)$$

Per capirne meglio l'importanza

(teorema della probabilità delle cause)

All'esame:

lo studente può essere
- *preparato* H_0
- *non-preparato* $H_1 \ (= \overline{H}_0)$



Evento: S = *lo studente supera l'esame*

$$S = (S \cap H_0) \cup (S \cap H_1)$$

$$P(S) = P(S \cap H_0) + P(S \cap H_1) =$$

$$= P(H_0) \cdot P(S \mid H_0) + P(H_1) \cdot P(S \mid H_1)$$

Il prof. "sa" che: $P(H_0) = 0.75$ (prob. che uno studente sia *preparato*)

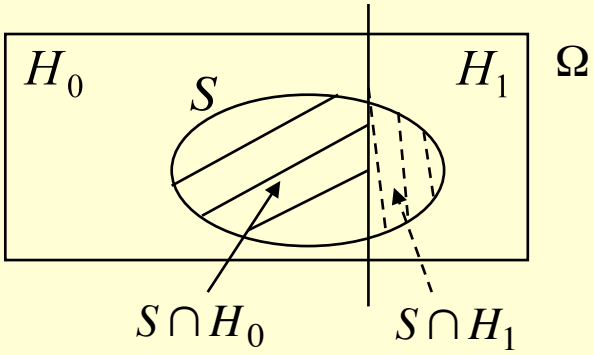
$P(H_1) = 0.25$ (prob. che uno studente sia *non - preparato*)

e che $P(S \mid H_0) = 0.85$ (prob. che, *essendo preparato*, superi l'esame)

$P(S \mid H_1) = 0.20$ (prob. che, *non essendo preparato*, superi l'esame)

... applicando il teorema

(teorema della probabilità delle cause)



Uno studente si presenta all'esame

$P(H_0) = 0.75$ (prob. che lo studente sia *preparato*)

$P(H_1) = 0.25$ (prob. che lo studente sia *non - preparato*)

$P(S \mid H_0) = 0.85$ (prob. che, *essendo preparato*, superi l'esame)

$P(S \mid H_1) = 0.20$ (prob. che, *non essendo preparato*, superi l'esame)

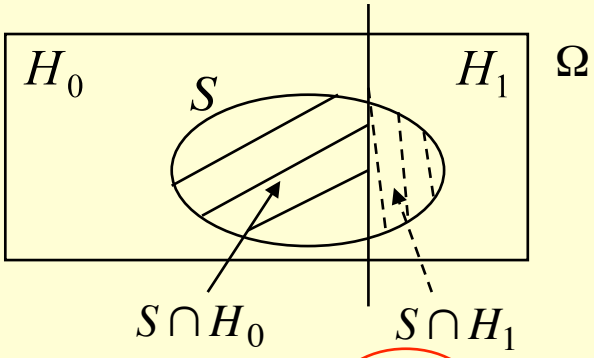
Lo studente supera l'esame

$$P(H_0 \mid S) = \frac{P(H_0) \cdot P(S \mid H_0)}{P(H_0) \cdot P(S \mid H_0) + P(H_1) \cdot P(S \mid H_1)} = 0.93$$

$$P(H_1 \mid S) = \frac{P(H_1) \cdot P(S \mid H_1)}{P(S)} = 0.07$$

... applicando il teorema

(teorema della probabilità delle cause)



Uno studente si presenta all'esame

$P(H_0) = 0.75$ (prob. che lo studente sia *preparato*)

$P(H_1) = 0.25$ (prob. che lo studente sia *non - preparato*)

$P(S \mid H_0) = 0.85$ (prob. che, *essendo preparato*, superi l'esame)

$P(S \mid H_1) = 0.20$ (prob. che, *non essendo preparato*, superi l'esame)

Lo studente supera l'esame

$$P(H_0 \mid S) = \frac{P(H_0) \cdot P(S \mid H_0)}{P(H_0) \cdot P(S \mid H_0) + P(H_1) \cdot P(S \mid H_1)} = 0.93$$

$$P(H_1 \mid S) = \frac{P(H_1) \cdot P(S \mid H_1)}{P(S)} = 0.07$$

Secondo tempo

Azione 2

L’indiano *Arturo* accetta di giocare. Smazza e trova una carta nera:

vince il cow boy *Piero*.

L’indiano Arturo si chiede: il cow boy ha barato?

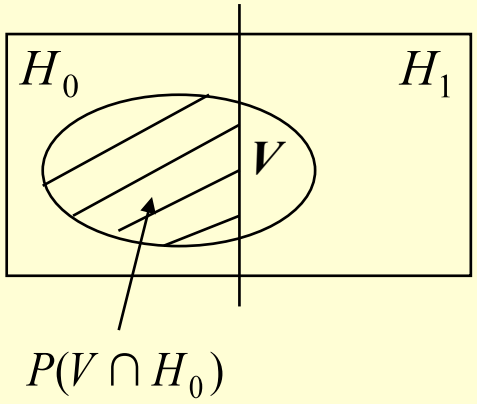
Interviene (*ma nel film non si vede!*) il reverendo T. Bayes con il suo teorema.

Per applicarlo, l’indiano Arturo ha bisogno dei numeri.

epílogo

Ipotesi

- *Piero è baro* H_0
- *Piero non è baro* $H_1 \quad (= \overline{H}_0)$



$$P(V \cap H_0) = P(H_0) \cdot P(V \mid H_0)$$
$$P(V \cap H_0) = P(V) \cdot P(H_0 \mid V)$$

Evento: V = la carta è nera ⇨ Piero vince

$$P(H_0 \mid V) = \frac{P(H_0) \cdot P(V \mid H_0)}{P(H_0) \cdot P(V \mid H_0) + P(H_1) \cdot P(V \mid H_1)} \quad \left(= \frac{P(H_0) \cdot P(V \mid H_0)}{P(V)} \right)$$

Probabilità che Piero abbia vinto barando

(che, avendo vinto, sia un baro)

$$P(H_0 \mid V) = \frac{P(H_0) \cdot P(V \mid H_0)}{P(H_0) \cdot P(V \mid H_0) + P(H_1) \cdot P(V \mid H_1)}$$

- Probabilità che Arturo attribuisce all'ipotesi (evento) H_0 che Piero sia baro:

$$P(H_0) = P(H_1) = \frac{1}{2} = 0.5$$ (*l'indiano è malfidato!*)

- Probabilità che Piero vinca senza barare: $P(V \mid H_1) = \frac{1}{2}$
- Probabilità che Piero vinca barando: $P(V \mid H_0) = \frac{3}{4}$ (ad esempio)
- Il calcolo

$$P(H_0 \mid V) = \frac{\frac{1}{2} \cdot \frac{3}{4}}{\frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{2}} = 0.6$$

$$P(H_0 \mid V) = \frac{P(H_0) \cdot P(V \mid H_0)}{P(H_0) \cdot P(V \mid H_0) + P(H_1) \cdot P(V \mid H_1)}$$

- Probabilità che Arturo attribuisce all'ipotesi (evento) H_0 che Piero sia baro:

$$P(H_0) = P(H_1) = \frac{1}{2} = 0.5$$ (*l'indiano è malfidato!*)

- Probabilità che Piero vinca senza barare: $P(V \mid H_1) = \frac{1}{2}$
- Probabilità che Piero vinca barando: $P(V \mid H_0) = \frac{3}{4}$ (ad esempio)
- Il calcolo

$$P(H_0 \mid V) = \frac{\frac{1}{2} \cdot \frac{3}{4}}{\frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{2}} = 0.6$$

FINE

Altre incertezze

- Altre incertezze sorgono da misurazioni che ci toccano molto più da vicino: le diagnosi mediche (non molto diverse da quelle che formula l'insegnante quando "visita" uno studente).
- Le diagnosi sono formulate a seguito di un test diagnostico che consiste, generalmente, in una misurazione.
- Attraverso il test viene rilevato un carattere quantitativo (spesso sintetizzato qualitativamente) e la diagnosi si esprime - sempre in generale - come un carattere qualitativo dicotomico (*sano/malato, affetto/non affetto, presente/assente*).
- Siamo quindi in ambito statistico e, ancora, statistica e probabilità forniscono misure dell'incertezza.

Un esempio paradigmatico con dati reali

Test di Test

test *statistico* per test *diagnostico*

(dove interviene la misura dell'incertezza della diagnosi)

Aneurisma dell'aorta addominale

- *Patologia*: "aneurisma dell'aorta addominale", consiste nella presenza nell'aorta di una dilatazione anomala.
- *Calibro dell'aorta in stato di salute*: 1.5 – 2.0 *cm*
- *Anomalia*: > 4 *cm*
- *Conseguenze*: rottura dell'aneurisma
- *Cura*: chirurgica
- *Soglia per intervenire chirurgicamente*: $k = 5$ *cm*

Nota:

è fondamentale rilevare quanto più precisamente possibile il calibro dell'aneurisma, per garantire un corretto intervento.

Il problema: validazione di un nuovo test diagnostico

- Test diagnostico di riferimento *(golden test)*:

  lo strumento di misura usuale si basa sulla *TAC (**T**omografia **A**ssiale **C**omputerizzata)*

- In seguito alla misura del calibro dell'aorta, i pazienti sono classificati in *rischio medio-basso* e *rischio alto* ($k \geq 5$), per semplicità *Sani* e *Malati*

- Nuovo strumento basato sugli ***ultrasuoni*** *(US)*: meno invasivo, più rapido, più economico

- *Problema*: validare il nuovo *test diagnostico* rispetto alla sua capacità di discriminare tra *Sani* e *Malati*

- *Metodologia*: analisi statistica dei dati → *test statistico*

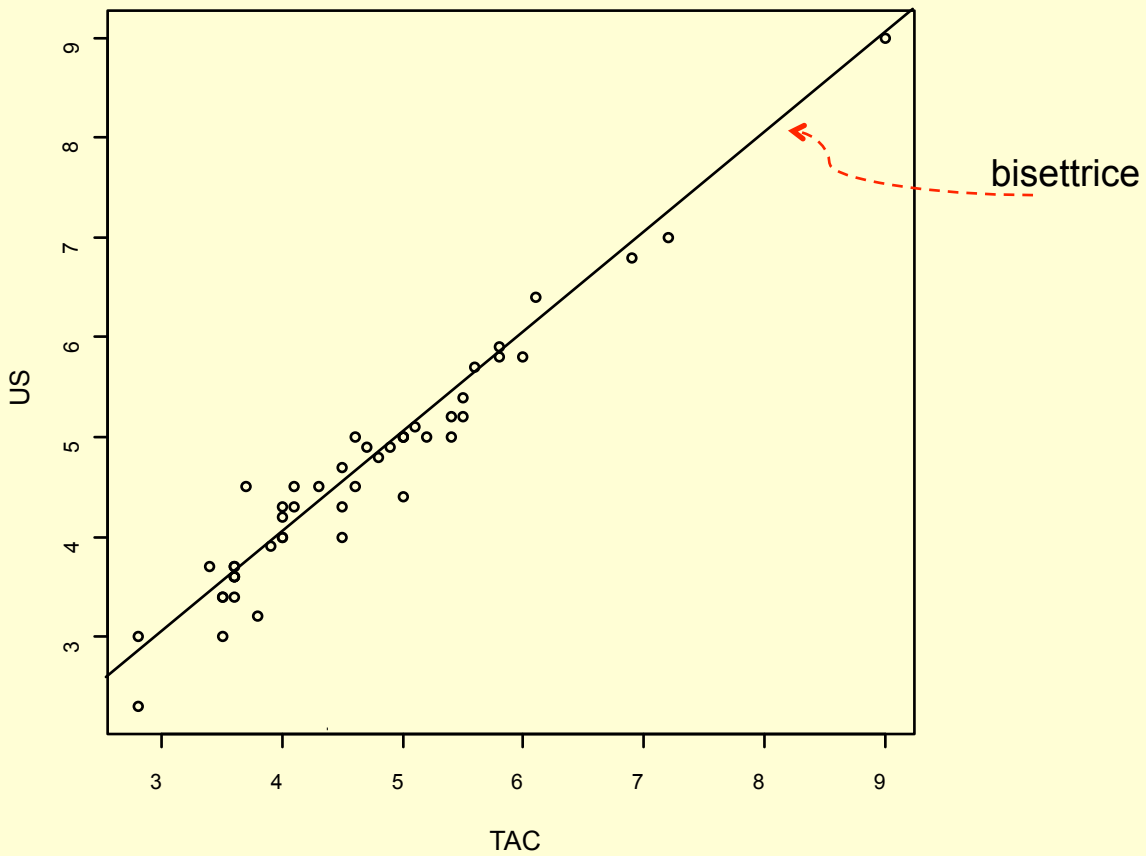
I dati reali *(calibri in mm)*

(cut-off k=5.0 cm)

TAC	US	TAC	US	TAC	US
35	30	50	50	34	37
58	59	54	52	36	36
55	54	35	34	45	40
37	45	28	23	49	49
54	50	50	50	61	64
58	58	40	42	36	37
45	47	36	36	28	30
41	43	43	45	40	40
46	50	56	57	36	34
40	40	36	37	38	32
60	58	50	44	51	51
52	50	72	70	47	49
40	43	55	52	69	68
45	43	41	45	36	36
90	90	50	50	39	39
48	48	35	34	46	45

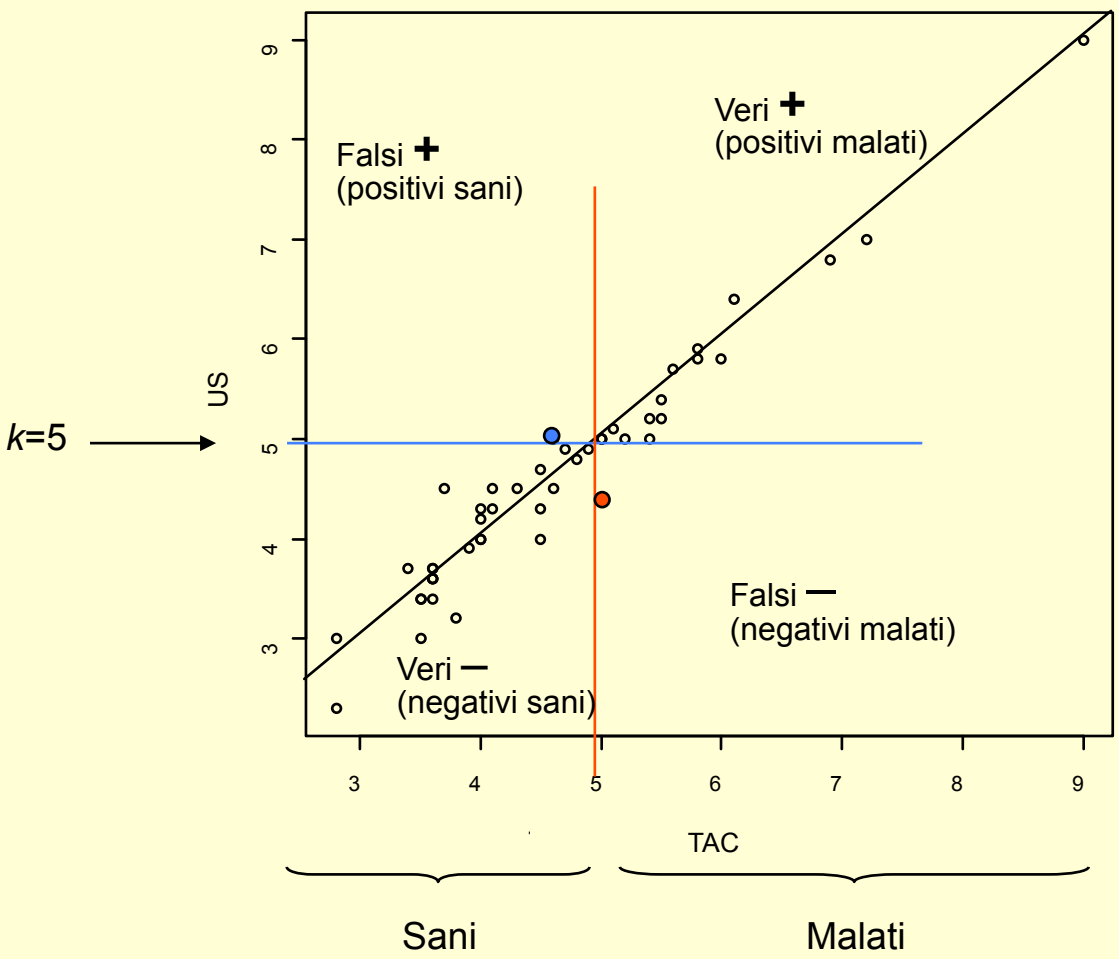
Una prima analisi esplorativa

diagramma dei calibri misurati con TAC e con US



Classificazione dei pazienti

in base al test diagnostico US



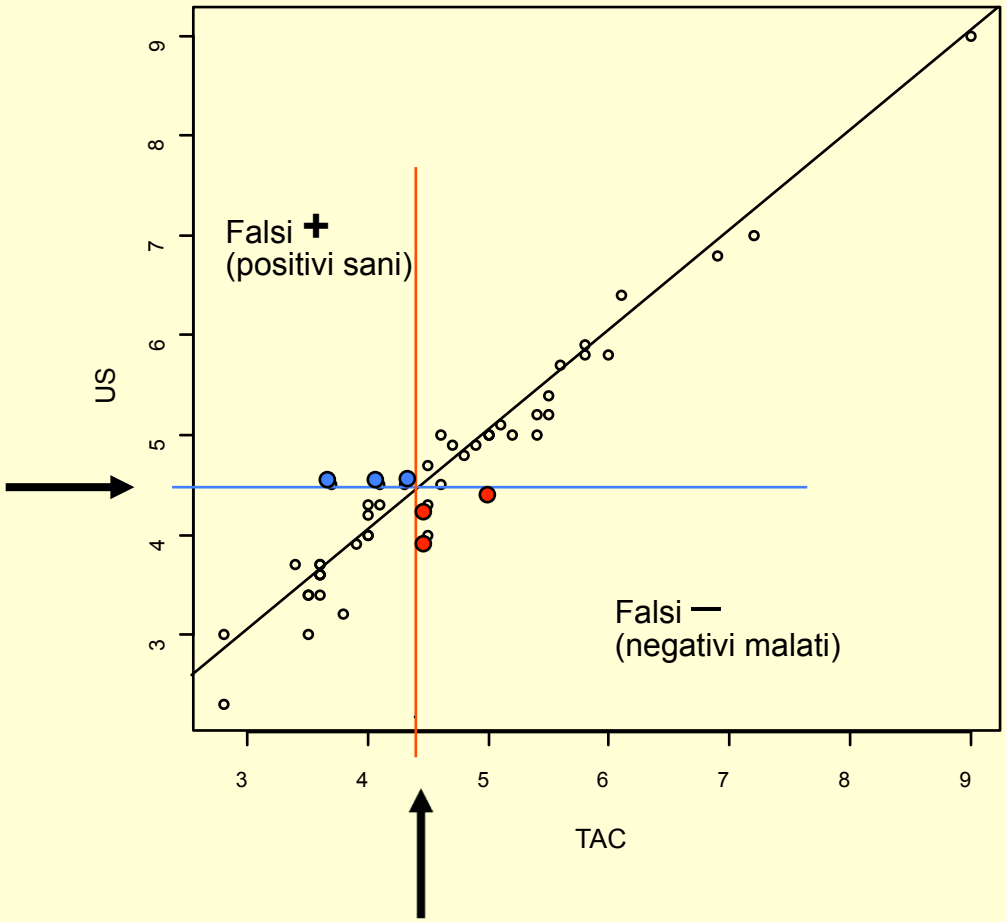
Conseguenze del criterio di valutazione

(cut-off k=4.5 cm)

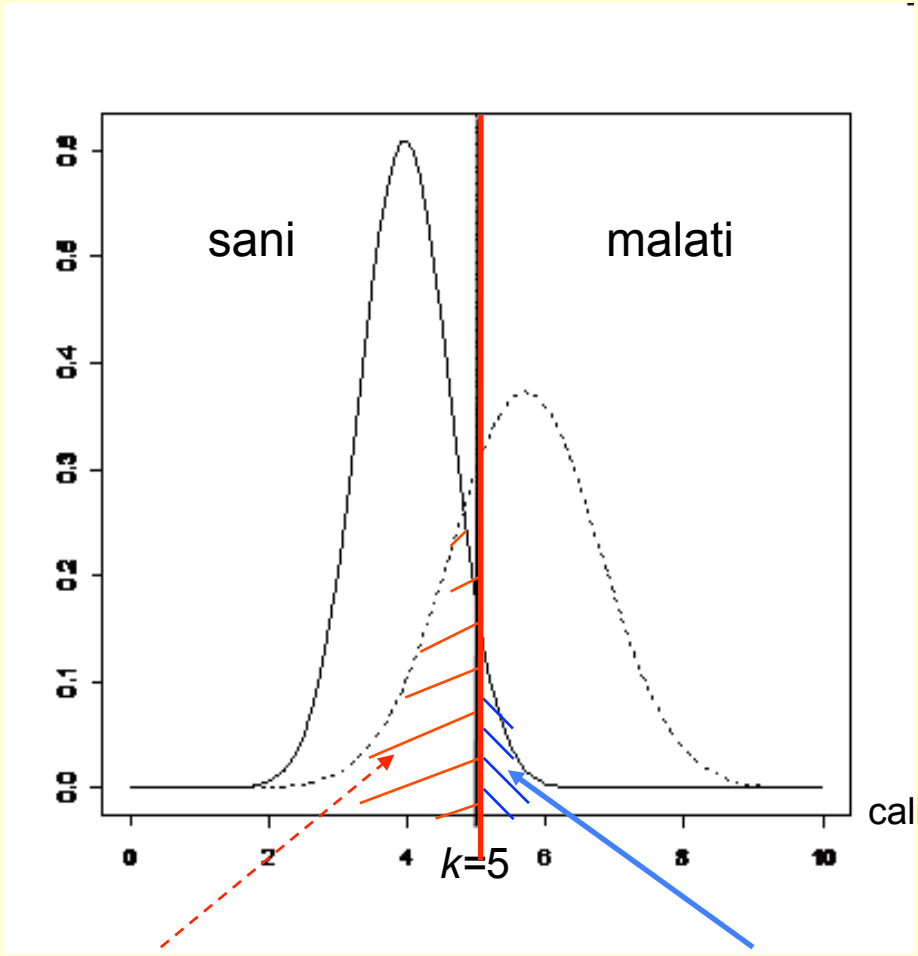
TAC	US	TAC	US	TAC	US
35	30	50	50	34	37
58	59	54	52	36	36
55	54	35	34	45	40
37	45	28	23	49	49
54	50	50	50	61	64
58	58	40	42	36	37
45	47	36	36	28	30
41	43	43	45	40	40
46	50	56	57	36	34
40	40	36	37	38	32
60	58	50	44	51	51
52	50	72	70	47	49
40	43	55	52	69	68
45	43	41	45	36	36
90	90	50	50	39	39
48	48	35	34	46	45

Classificazione dei pazienti

(cut-off k=4.5 cm)



Rappresentazione delle distribuzioni (*normali*) dei calibri misurati con *US* nei Sani e Malati



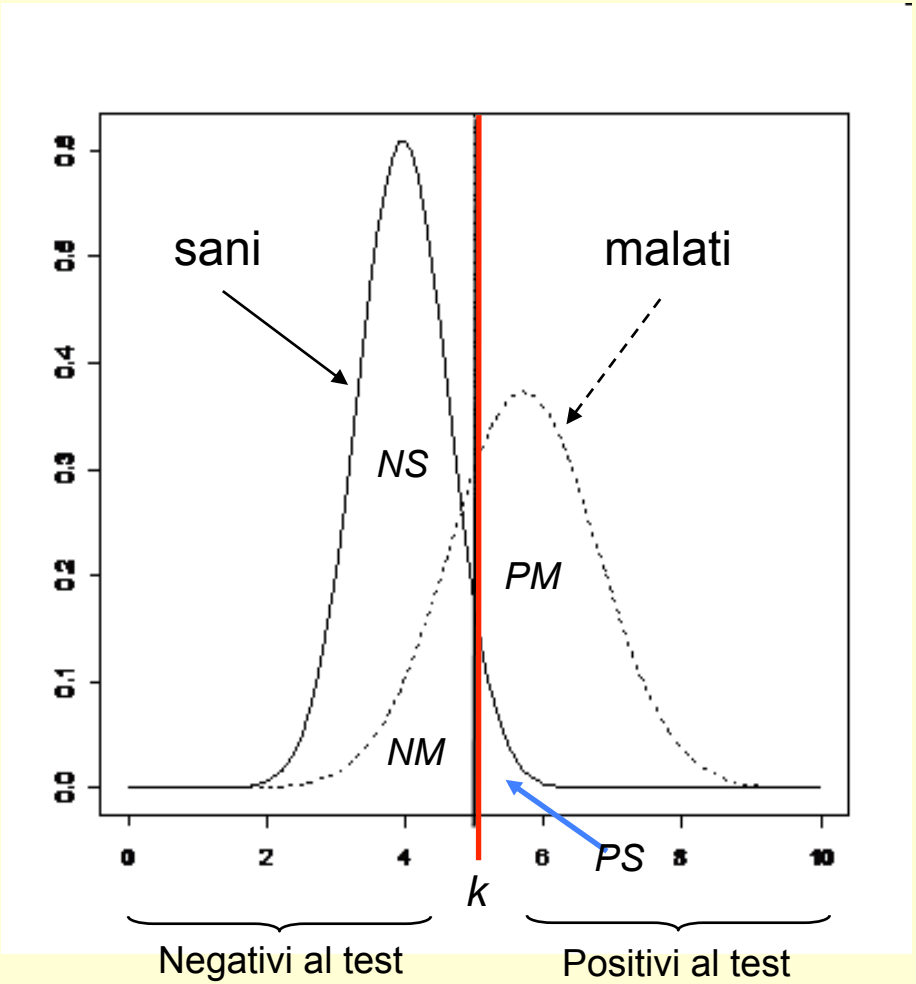
Falsi positivi e Falsi Negativi
***P**ositivi **S**ani* (*PS*) e ***N**egativi **M**alati* (*NM*)

- In genere le distribuzioni dei pazienti SANI e dei pazienti MALATI sottoposti a un test diagnostico si sovrappongono parzialmente dando luogo a *aree di confondimento*.

- Per ogni soglia *k* (*criterion value* o *cut-off*) ci saranno pazienti:

1. *Malati* correttamente classificati come *Positivi* → (*PM* = positivi malati)
2. *Malati* classificati come *Negativi* → (*NM* = negativi malati)
3. *Sani* correttamente classificati come *Negativi* → (*NS* = negativi sani)
4. *Sani* classificati come *Positivi* → (*PS* = positivi sani)

Riepilogo grafico



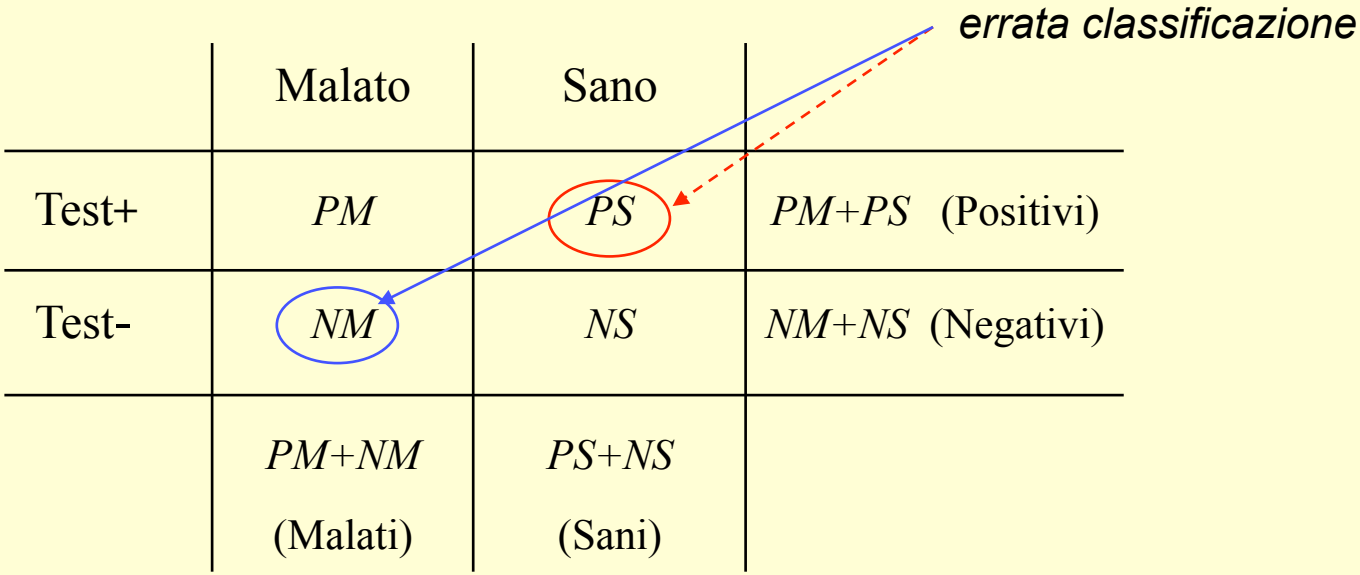
La tabella a doppia entrata = *matrice di confusione*

(confounding matrix)

	Malato	Sano	
Test+	*PM*	*PS*	*PM*+*PS* (Positivi)
Test-	*NM*	*NS*	*NM*+*NS* (Negativi)
	PM+*NM* (Malati)	*PS*+*NS* (Sani)	

La tabella a doppia entrata = *matrice di confusione*

(confounding matrix)



La validità del test diagnostico può essere misurata in base alle proporzioni di Falsi + e Falsi – , quanto più basse sono tanto più valido sarà il test.
In altri termini: in base alla capacità di corretta classificazione. Ma …

… ma **la realtà è quella che è**!

Che cosa possiamo chiedere a un test diagnostico?

- Che sia ***accurato***: alta capacità di corretta classificazione
  (elevata proporzione di *PM* e *NS* rispetto al totale dei pazienti osservati)

- Che sia ***sensibile*** alla malattia: alta capacità di classificare i *Malat*i come *Positivi* al test
  (elevata proporzione di *PM* rispetto al *totale dei Malati*)

  Nota: un test è sensibile al 100% quando **tutti** i *Malati* risultano *Positivi.*

- Che sia ***specifico***: alta capacità di classificare i *Sani* come *Negativi* al test
  (elevata proporzione di *NS* rispetto al *totale dei Sani*).

  Nota: un test è specifico al 100% quando tutti i *Sani* risultano *Negativi.*

Un test sensibile e specifico al 100% non lascerebbe dubbi !

	Malato	Sano	
Test +	*PM*	*PS*	*PM+PS*
Test -	*NM*	*NS*	*NM+NS*
	PM+NM	*PS+NS*	

$$accuratezza = \frac{PM + NS}{PM + PS + NM + NS}$$

in termini predittivi: *prob. di corretta class.*

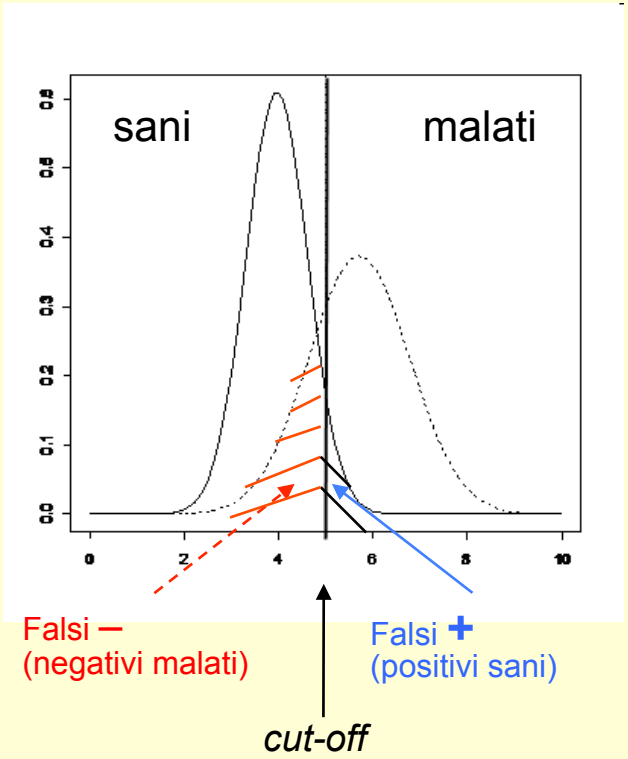
$$sensibilità = \frac{PM}{PM + NM} \qquad \left(Prob(P \mid M)\right)$$

$$specificità = \frac{NS}{NS + PS} \qquad \left(Prob(N \mid S)\right)$$

Sensibilità e *specificità* dipendono dal *cut-off*

L'aspirazione è di minimizzare gli errori di classificazione:

ridurre il rischio di *Falsi*+ e *Falsi*–



Soglia alta

Test poco sensibile
Elevata specificità

- sottostima la prop. di *Malati*
- basso rischio di Falsi +
- più Falsi –
- "protegge" i *Sani*

Soglia bassa

Test molto sensibile
Scarsa specificità

- sottostima la prop. di *Sani*
- basso rischio Falsi –
- più Falsi +
- individua più *Malati*

Che fare? Come fissare la soglia?

Elevata *sensibilità* e bassa *specificità* o viceversa?

Dipende dall'obiettivo del test (e dal contesto clinico-epidemiologico)

- Malattia a grave rischio, prevenibile con intervento immediato:

⇨ test molto **sensibile** alla malattia, seppure poco specifico, per non rischiare di perdere dei *Malati* (anche a discapito di doverne "spaventare" alcuni di più!).

- Malattia con conseguenze non gravi, terapie con effetti collaterali rischiosi, scarsità di risorse (umane, finanziarie, farmacologiche), necessità di diagnosi prudenziale:

⇨ test molto **specifico** anche a discapito della sensibilità, meno Falsi+ con maggiore rischio di Falsi –

… e dunque?

(interviene ancora la Probabilità !)

- Il caso di studio (dati aneurisma dell'aorta addominale)
+ Due esempi fittizi (1), (2)

	Malato	Sano	
Test +	17	1	18
Test -	1	29	30
	18	30	48

$$\text{sensibilità} = \frac{17}{18} = 0.94$$

$$\text{specificità} = \frac{29}{30} = 0.97$$

$$\text{accuratezza} = \frac{17+29}{48} = 0.96$$

(1)

	Malato	Sano	
Test +	85	180	265
Test -	15	720	735
	100	900	1000

$\text{sensibilità} = 0.85$
$\text{specificità} = 0.80$
$\text{accur.} = 0.81$

(2)

	Malato	Sano	
Test +	680	40	720
Test -	120	160	280
	800	200	1000

$\text{sensibilità} = 0.85$
$\text{specificità} = 0.80$
$\text{accur.} = 0.84$

Ma l'incidenza (**prevalenza**) della malattia nelle due popolazioni è ben diversa!

Prevalenza e teorema di Bayes

- ***Prevalenza*** di malattia: $\frac{PM + NM}{n\ \ totale} = \frac{totale\ \ malati}{n\ \ totale} \quad (\ Prob(M)\ )$

Caso (1) $Prevalenza = \frac{100}{1000} = 10\%$; Caso (2) $Prevalenza = \frac{800}{1000} = 80\%$

... e finalmente

- Diagnosi medica: probabilità che un *Positivo* sia *Malato*

$$Prob(M \mid Positivo) = \frac{P(M) \cdot P(Pos \mid M)}{P(Pos)} = \frac{P(M) \cdot P(Pos \mid M)}{P(M) \cdot P(Pos \mid M) + P(S) \cdot P(Pos \mid S)}$$

Nell'esempio (fittizio) considerato:

(1) $Prevalenza = 10\%$

$$Prob(M \mid Positivo) = \frac{85}{265} = 32.07\%$$

(2) $Prevalenza = 80\%$

$$Prob(M \mid Positivo) = \frac{680}{720} = 94.44\%$$

In ambulatorio

- La mamma porta dal pediatra il bambino di 8 anni febbricitante, con la pelle visibilmente cosparsa di puntini rossi e la lingua violacea:

  - "dottore, sarà *scarlattina*?"

- Il medico passa un *tampone faringeo* sulle tonsille del bambino, quindi lo inserisce nell'apposito astuccio con i reagenti e attende l'esito.
- Il *test diagnostico* risulta *positivo* (il tampone rivela la presenza batterica).
- Il medico sa che nella popolazione in età 6 -12 anni la proporzione di bambini che si ammalano di scarlattina è del 10% (*prevalenza=proporzione di soggetti affetti da una certa malattia*) . Sa anche che il tampone fornisce la risposta corretta (*veri positivi* e *veri negativi*) nel 98% dei casi.
- Sulla base di questi dati e dell'esito del test diagnostico, il medico formula la diagnosi dicendo alla mamma quanto vale la probabilità che il bambino abbia davvero la scarlattina.

Per arrivare alla diagnosi abbiamo bisogno degli opportuni strumenti statistici

Formalizziamo il problema *(definiamo gli eventi e le probabilità)*

- *M* = il bimbo ha la scarlattina: *Malato*
- *S* = il bimbo NON ha la scarlattina: *Sano*
- *P* = il test *(tampone faringeo)* risulta *Positivo*
- *N* = il test risulta *Negativo*

Il medico sa *(... in base alla propria esperienza ...)*:

$$Prob(M) = 0.10 \rightarrow Prob(S) = 0.90 \quad (probabilità\ a\ priori)$$
$$Prob(P \mid M) = Prob(N \mid S) = 0.98$$
$$Prob(N \mid M) = Prob(P \mid S) = 0.02$$

Il medico esamina il tampone faringeo che risulta *Positivo* (esito del test):

→ alla luce del risultato, come ragiona per formulare la diagnosi?

Il medico applica il *teorema di Bayes*

In altri termini: alla luce dell'esito del test ***aggiorna*** la sua probabilità *a priori*,

$$Prob(M) \rightarrow Prob(M \mid P) \quad (probabilità\ a\ posteriori)$$

$$Prob(M \mid P) = \frac{Prob(M) \cdot Prob(P \mid M)}{Prob(M) \cdot Prob(P \mid M) + Prob(S) \cdot Prob(P \mid S)}$$

$$= \frac{0.10 \cdot 0.98}{0.10 \cdot 0.98 + 0.90 \cdot 0.02} = 0.845$$

Visto l'esito del tampone faringeo *(evidenza sperimentale)* la valutazione della probabilità che il bimbo *abbia la scarlattina* è passata dal 10% all'85% (circa).

Attenzione però agli usi distorti *(stupefacenti!)*

della

Statistica


icocche
più del
100%
di frutta
biologico-organic
senza pectina aggiunta


più del
100%
di frutta
senza pectina aggiunta

“L’incertezza domina ovunque.

Tutta la nostra vita è immersa nell’incertezza;

nulla - all’infuori di ciò - si può affermare con certezza.”

(Bruno de Finetti, 1906-1985)