

Basi di statistica per gli avversari alla matematica

**SE NON SI CAPISCE
INTERROMPETEMI!**

Michele Rienzner, Ph.D.

Analista dati – Data scientist

Laurea in ingegneria per l'ambiente e il territorio, dottore di ricerca,
postdoc. Tecnico laureato - elaborazione dati presso l'Università di
Milano

La statistica

La matematica permette di trattare cose certe, ... ma cosa c'è di certo?

La statistica permette di maneggiare cose **incerte** e **valori variabili**.

La propaganda e la manipolazione giocano molto sull'incapacità della gente di considerare la probabilità degli eventi e puntano a fissare l'attenzione su eventi estremi, non impossibili, ma così rari da essere di scarso rilievo.

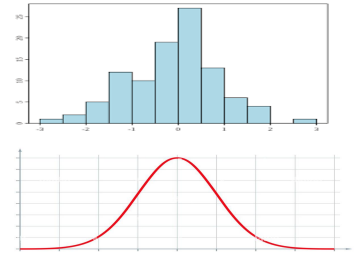
Obbligo di casco per non morire colpiti da un meteorite



Per smascherarli occorre capirci un po' di statistica.

Scaletta

1. Cosa si intende per «probabilità»
2. Come descrivere valori multipli: media e deviazione standard
3. L'istogramma di frequenza
4. La distribuzione di probabilità, la «NORMALE» e altre distribuzioni
5. Le sorgenti di errore e incertezza secondo il metodo di misura adottato
6. Cosa si intende in statistica per «**ERRORE**»
7. La variabilità dell'errore e il «**BIAS**»
8. L'**intervallo di confidenza** e la «**significatività statistica**»
9. Il **test** statistici di verifica delle ipotesi e il «**p-value**»
10. Un esempio di bias di stima e le sue conseguenze



La probabilità

Quantifica il grado di certezza che una cosa accada.

Si può esprimere come:

- numero tra 0 e 1 (0,25)
- frazione (1/4)
- percentuale (25%)

Probabilità 0 ➔ evento impossibile

Probabilità 1 ➔ evento certo

Probabilità vicine a 0 rendono l'evento raro (vincere al lotto) ma resta possibile per quanto piccola possa essere la possibilità.

Una probabilità vicine al 100% rendono l'evento quasi certo, ma esiste comunque la possibilità che non si verifichi.



La probabilità

Grado di certezza che una cosa accada.

Probabilità teorica

Si ottiene attraverso formule, è molto precisa a patto che:

- esista una formula per il caso in esame e
- il caso in esame rispecchi esattamente quello teorico su cui si basa la formula. Piccole differenze possono portare risultati totalmente fuorvianti.



Esempio

Probabilità teorica di fare 1 con un dado da 6 è $1/6$ (16,666%)

È esatta se tutti i numeri hanno la stessa probabilità di uscire.

La probabilità

Grado di certezza che una cosa accada.

Probabilità frequentista

Quante volte l'evento accade rispetto al totale dei tentativi.

Pro e contro:

- la misura è incerta, diventa più precisa all'aumentare dei tentativi,
- se il numero delle prove è basso, la stima della probabilità è molto incerta,
- si può applicare a tutto ciò che si può testare.



Esempio

Tiro **400** volte il dado e, ad esempio, ottengo 1 per **86** volte.

La probabilità è: $86 / 400 = 0,215 \rightarrow 22 \%$

La probabilità

Grado di certezza che una cosa accada.



Chi ha ragione, probabilità teorica (16,66%) o la frequentista (22%)?

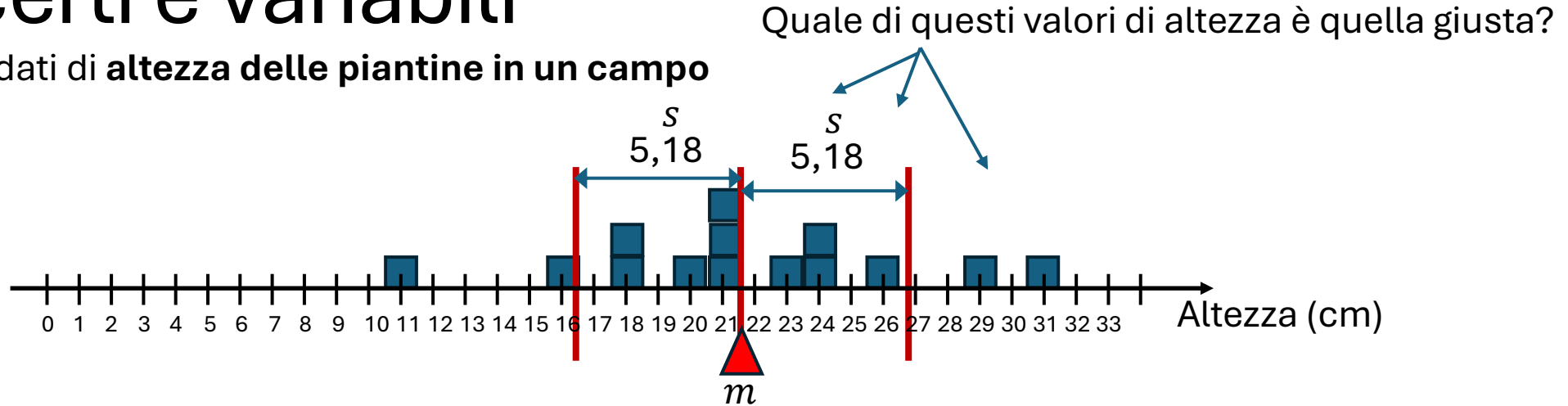


Se i tentativi sono sufficienti ha ragione la frequentista.

Il dado non non è bilanciato, ad sx. il lato dell'uno può essere più pesante per problemi termici durante la colata della resina (ho provato davvero con i miei dadi e ora ho capito perché perdo a Warhammer 40k... 🙄)

Dati incerti e variabili

Raccolgo molti dati di **altezza delle piantine in un campo**



Ciascuna è giusta ma sono troppi numeri! Devo «distillarli» in pochi numeri molto significativi:

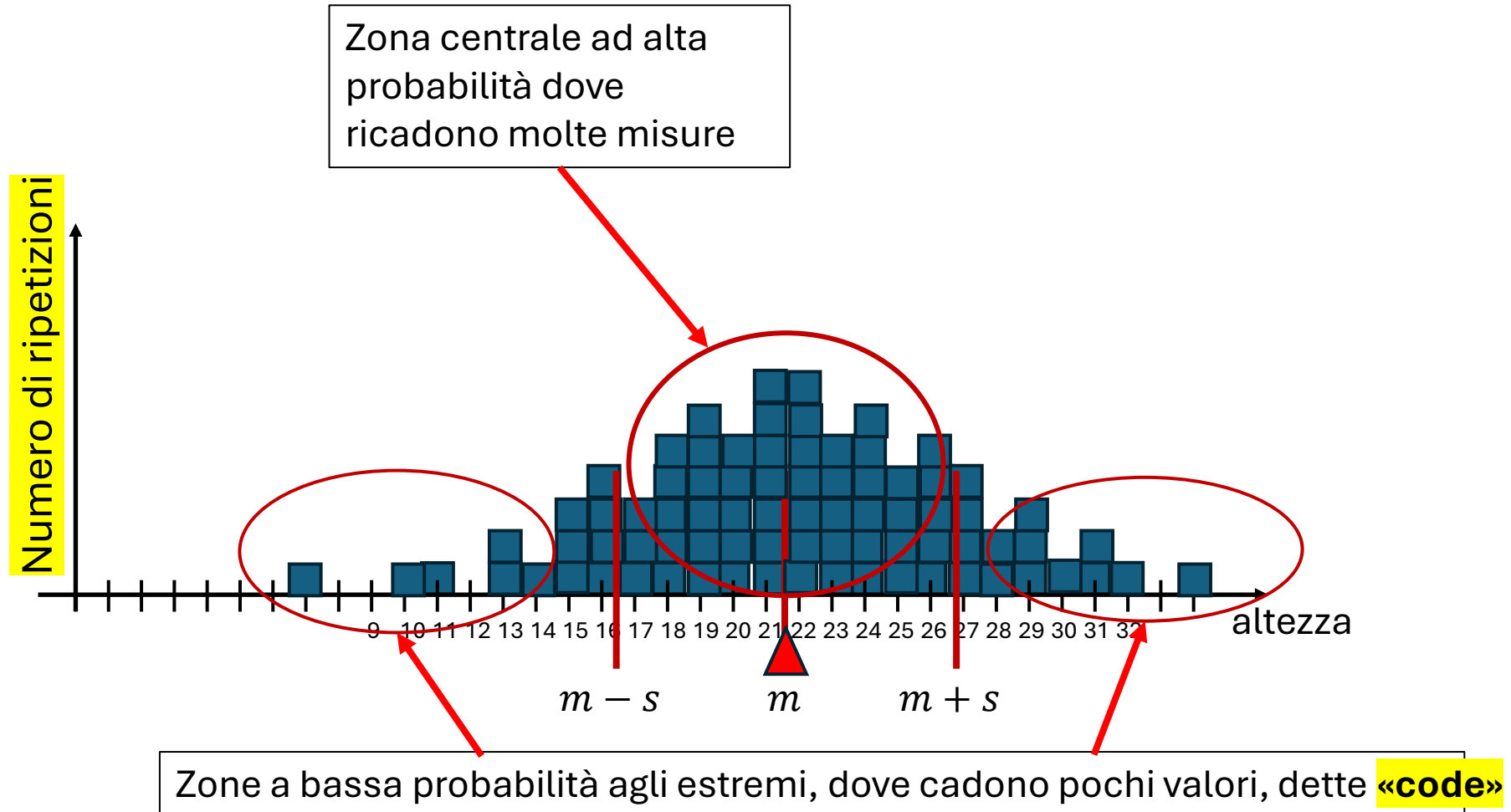
Media (media aritmetica) = dove sta il **punto centrale** dell'insieme di valori?

$$m = \frac{11 + 16 + 18 + 18 + 20 + 21 + 21 + 21 + 23 + 24 + 24 + 26 + 29 + 31}{14} = 21,64 \text{ cm}$$

Deviazione standard = quanto i valori si allargano attorno alla media? Quanto sono dispersi i dati?

$$s = \sqrt{\frac{(11-m)^2 + (16-m)^2 + (18-m)^2 + \dots}{14-1}} = 5,18 \text{ cm}$$

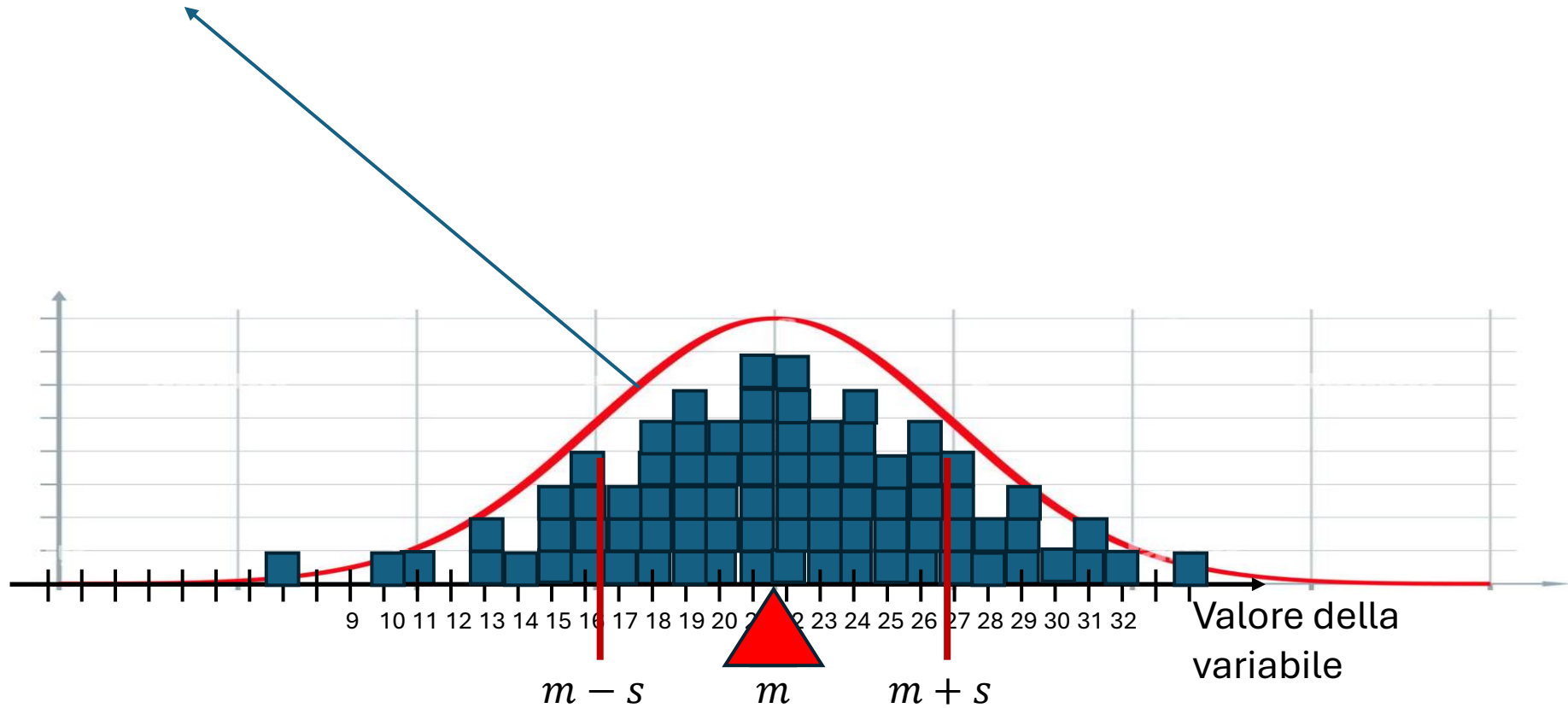
Istogramma di frequenza



La curva «normale»

È così comune incontrarla che viene solitamente
chiamata curva **«NORMALE»**

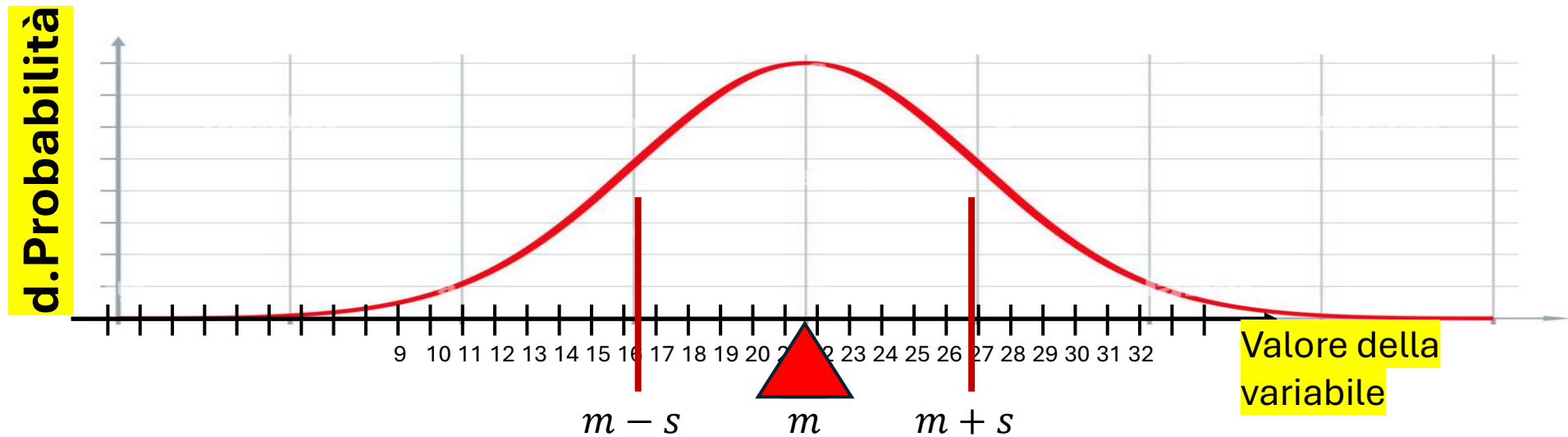
È una curva teorica a cui molti istogrammi somigliano



La curva «normale»

È così comune incontrarla che viene solitamente chiamata curva **«NORMALE»**

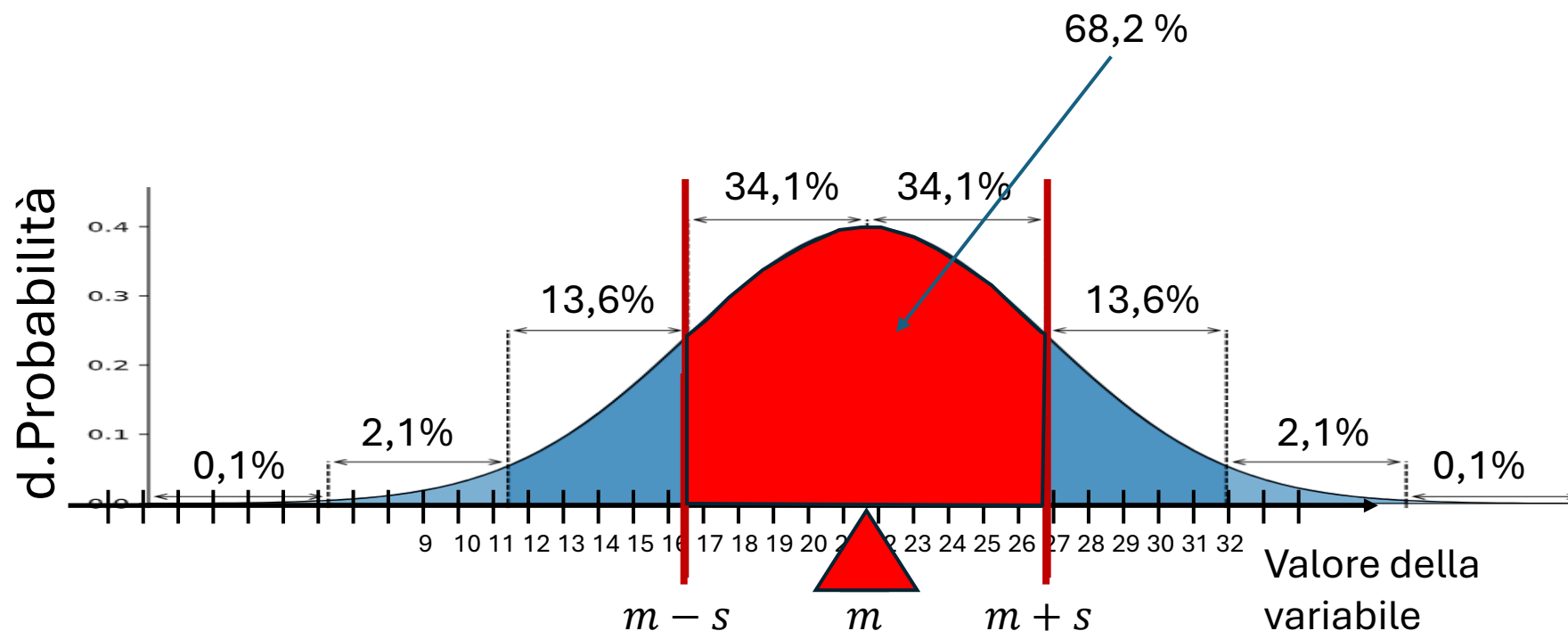
È una curva teorica a cui molti istogrammi somigliano



La curva «normale»

Come si usa la curva:

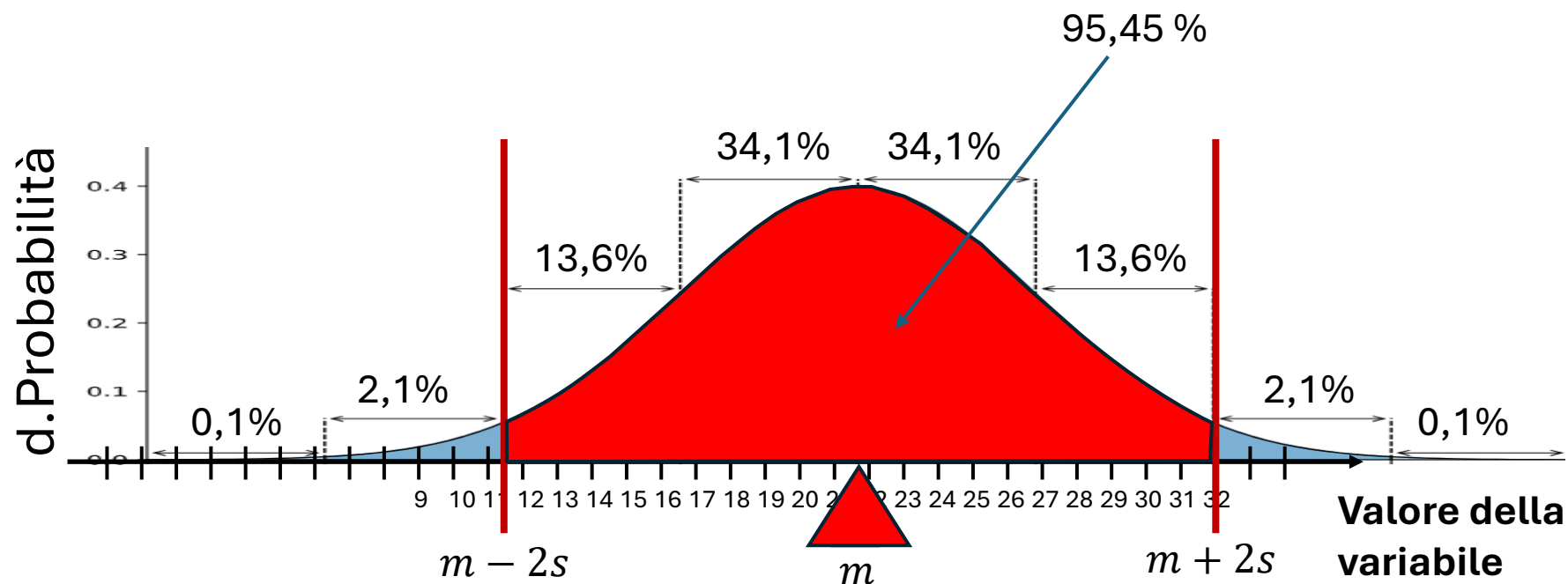
Per un intervallo di valori sulle X, ad es. da $m-s$ ad $m+s$,
l'area della curva nell'intervallo è la probabilità che il
valore estratto dalla variabile (ad es la prossima piantina
che misuro) cada entro quell'intervallo.



La curva «normale»

Come si usa la curva:

Per un intervallo di valori sulle X, ad es. da $m-s$ ad $m+s$,
l'area della curva nell'intervallo è la probabilità che il
valore estratto dalla variabile (ad es la prossima piantina
che misuro) cada entro quell'intervallo.

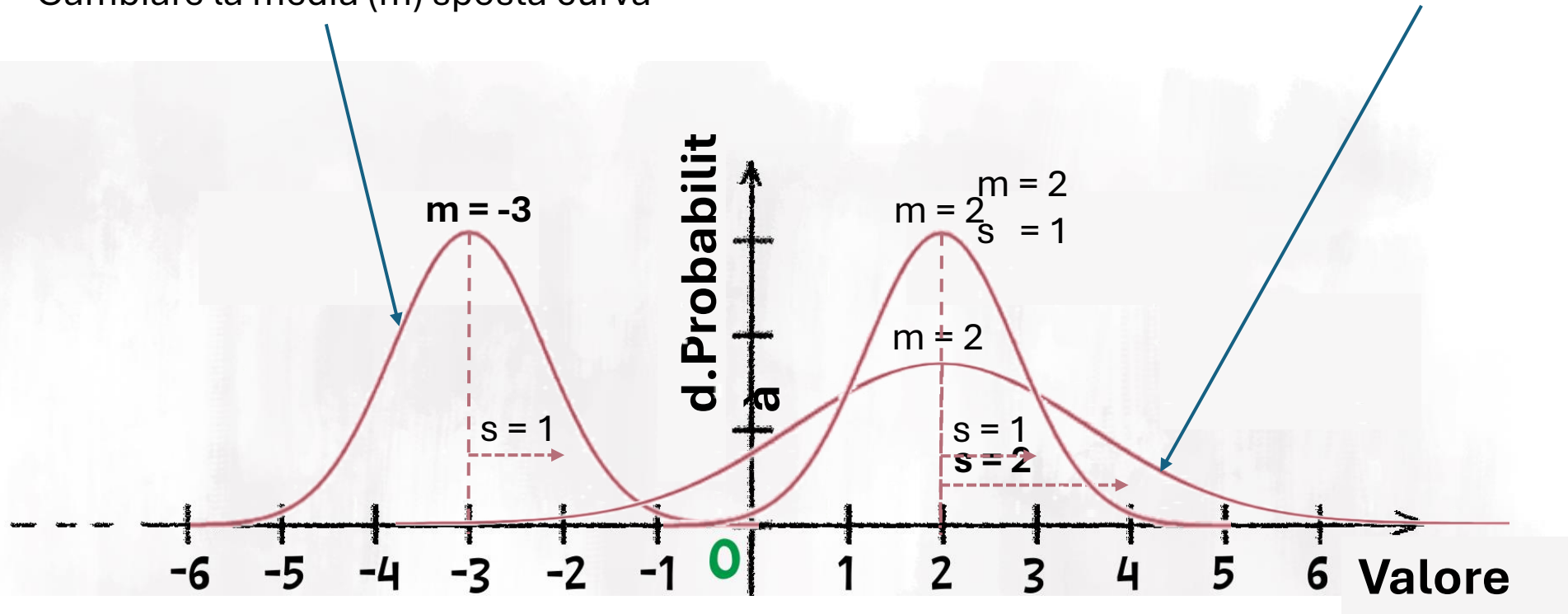


La curva «normale»

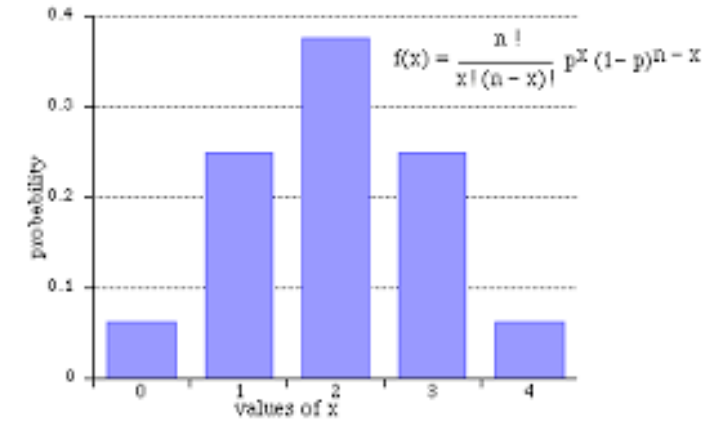
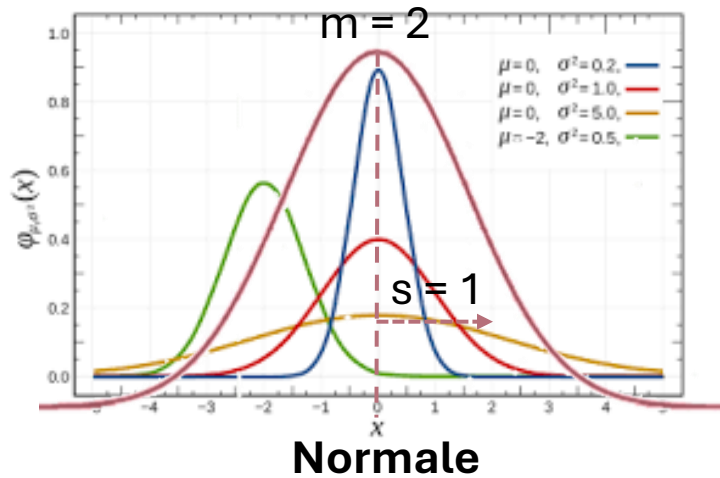
Cosa succede alla normale se cambio media e deviazione?

Cambiare la media (m) sposta curva

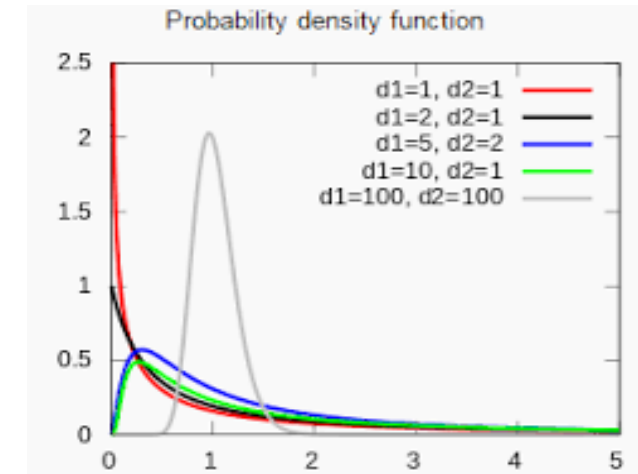
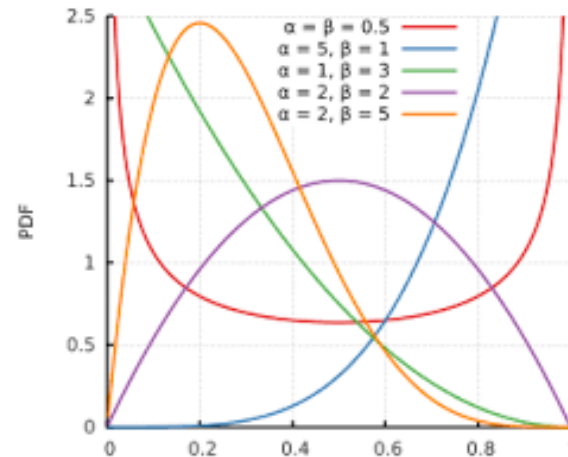
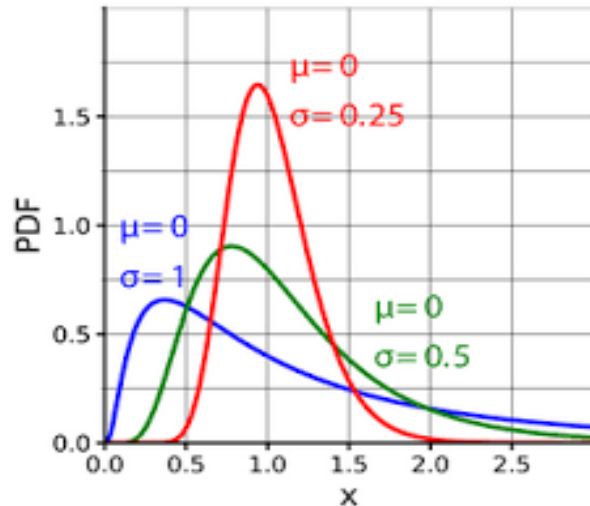
Aumentare la Deviazione standard (s) riduce la probabilità del picco e aumenta quella delle code. Allarga e abbassa.



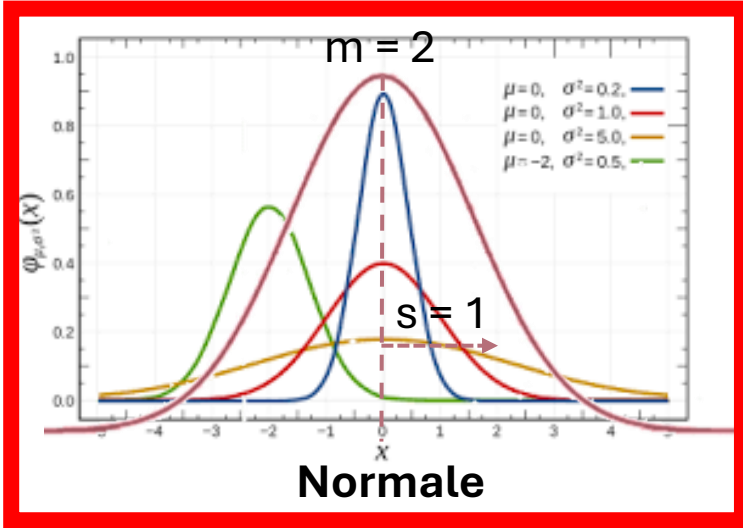
Distribuzioni di probabilità



Ci sono molte decine di distribuzioni, ciascuna ha il suo utilizzo.



Distribuzioni di probabilità



Ma la normale è special



È così comune incontrarla perché i valori che risultano dalla **somma di tanti influssi casuali** danno un istogramma che segue la distribuzione normale

Per questo è la distribuzione più usata, la più studiata e quella per cui ci sono più applicazioni e **test** per la verifica delle ipotesi

Misura diretta, indiretta e stima

Le sorgenti di incertezza nelle misure



Misura diretta – Misuro la **lunghezza** del tavolo confrontandolo con la **lunghezza** del metro.

Cause d'errore:

- esecuzione maldestra, strumento rotto, errori di lettura e di scrittura ← **errori grossolani**

Misura indiretta – si misura qualcosa che DOVREBBE essere legato alla cosa da misurare.



Ad esempio il “tampone rapido covid” dovrebbe misurare la presenza di una sufficiente quantità di antigene (proteina spike) nel muco nasale.

In realtà si osserva la variazione di colore di un reagente che reagisce all'antigene SE

- si evita esecuzione maldestra, strumento rotto, errori di lettura e di scrittura ← **errori grossolani**
- ci sono adeguate condizioni di temperatura, pH, ecc... ← **condizioni di validità**
- non ci sono altre sostanze o antigeni che diano falsi positivi ← **incertezze di misura**

Le misure indirette hanno quindi più cause d'errore, alcune controllabili altre no. E tutte potenzialmente possono andare a falsare l'esito del test producendo falsi positivi o falsi negativi.

Misura diretta, indiretta e stima

Le sorgenti di incertezza nelle misure

Stima - è una misura indiretta ottenuta attraverso qualche tipo di approccio statistico.

La stima è basata su solitamente su un «modello statistico» e il valore risultante è soggetto ad errore statistico che dipende da:

- errori accidentali informatici o di battitura
- errori nelle informazioni (dati in ingresso)
- semplificazioni del modello



← **errori grossolani**

← **incertezze** (simili alle misure indirette)

Inoltre il modello adottato è valido solo se:

- i dati rispettano le condizioni di validità del modello
- il caso esaminato combacia esattamente con quello ipotizzato nello sviluppo delle equazioni del modello.

← **condizioni di validità** (simili alle misure indirette)

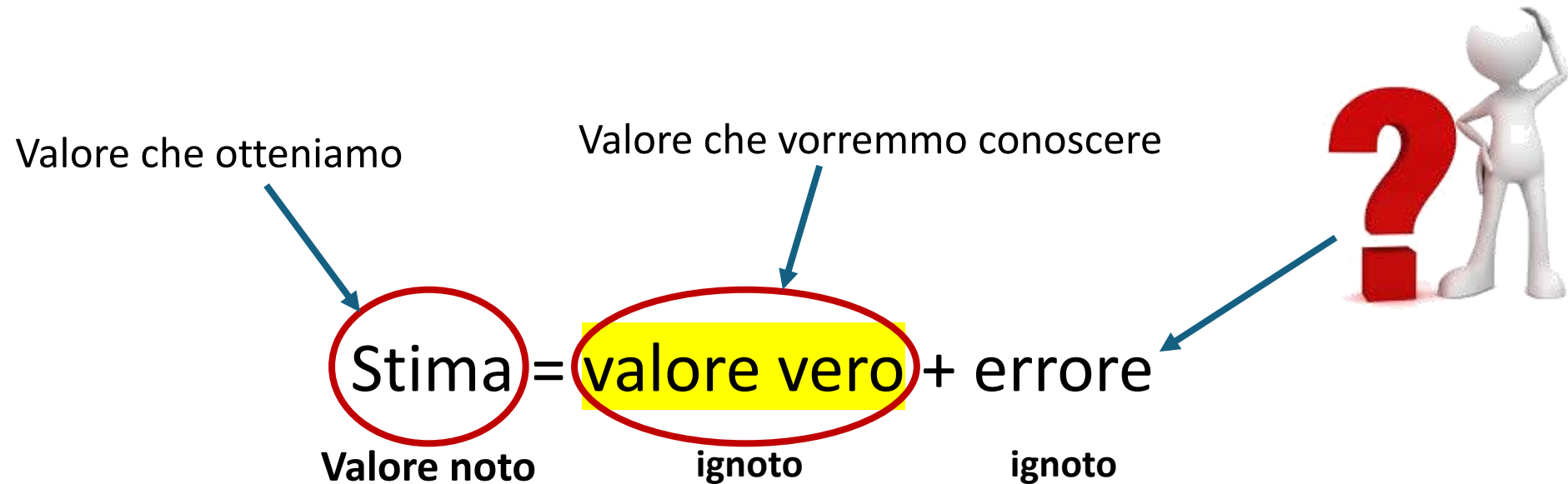
← **ambito di applicazione**

PRO: si possono ottenere stime di cose che altrimenti è impossibile misurare direttamente o indirettamente.

CONTRO: occorre prestare sempre molta attenzione a che tutte le condizioni siano valide o, anche in buona fede, vengono fuori valori totalmente fuorvianti con poche possibilità di riscontro.

L'ERRORE

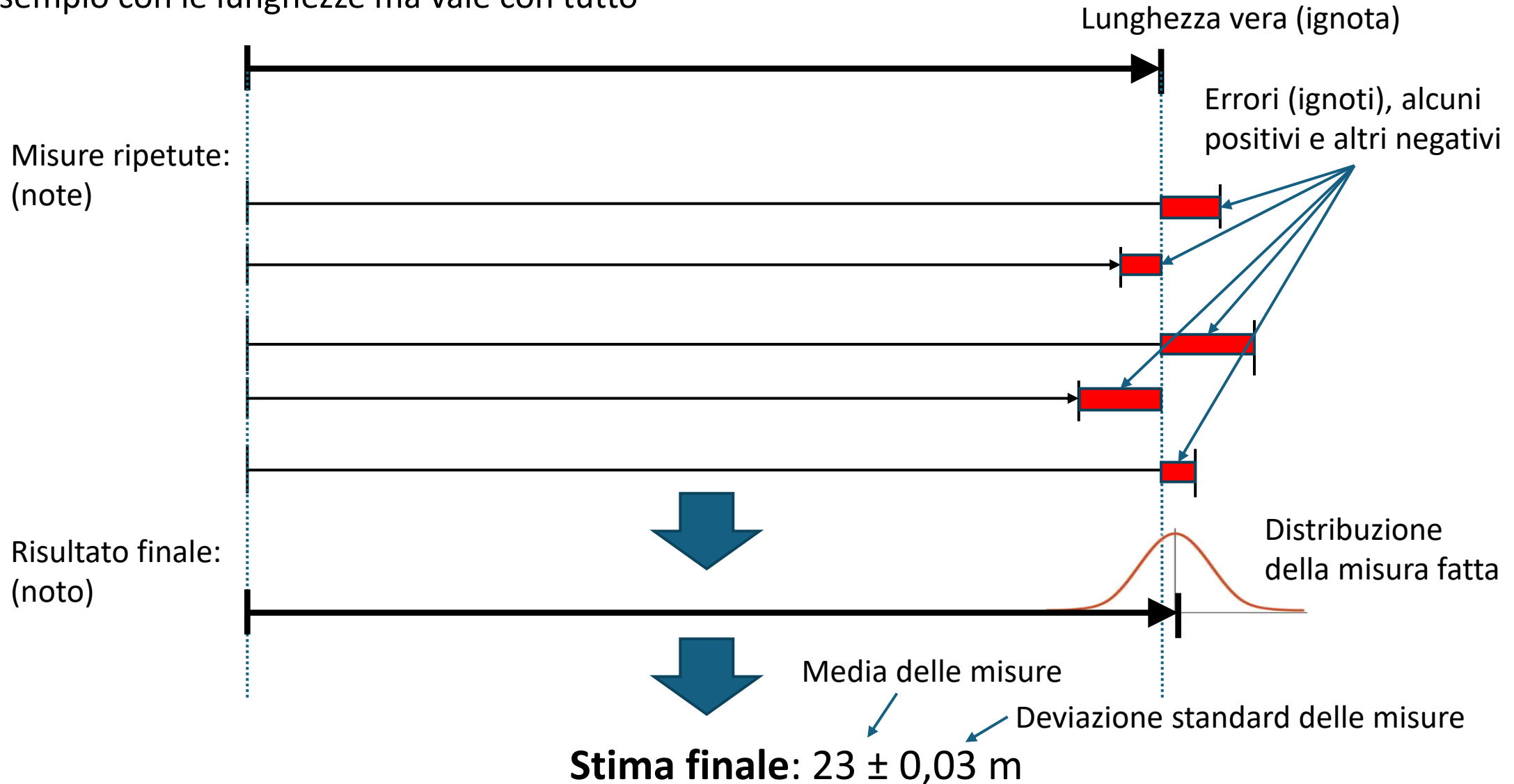
L'errore è la quantità che, durante la misurazione o la stima, viene aggiunta al valore vero a causa delle varie imprecisioni di ogni tipo.



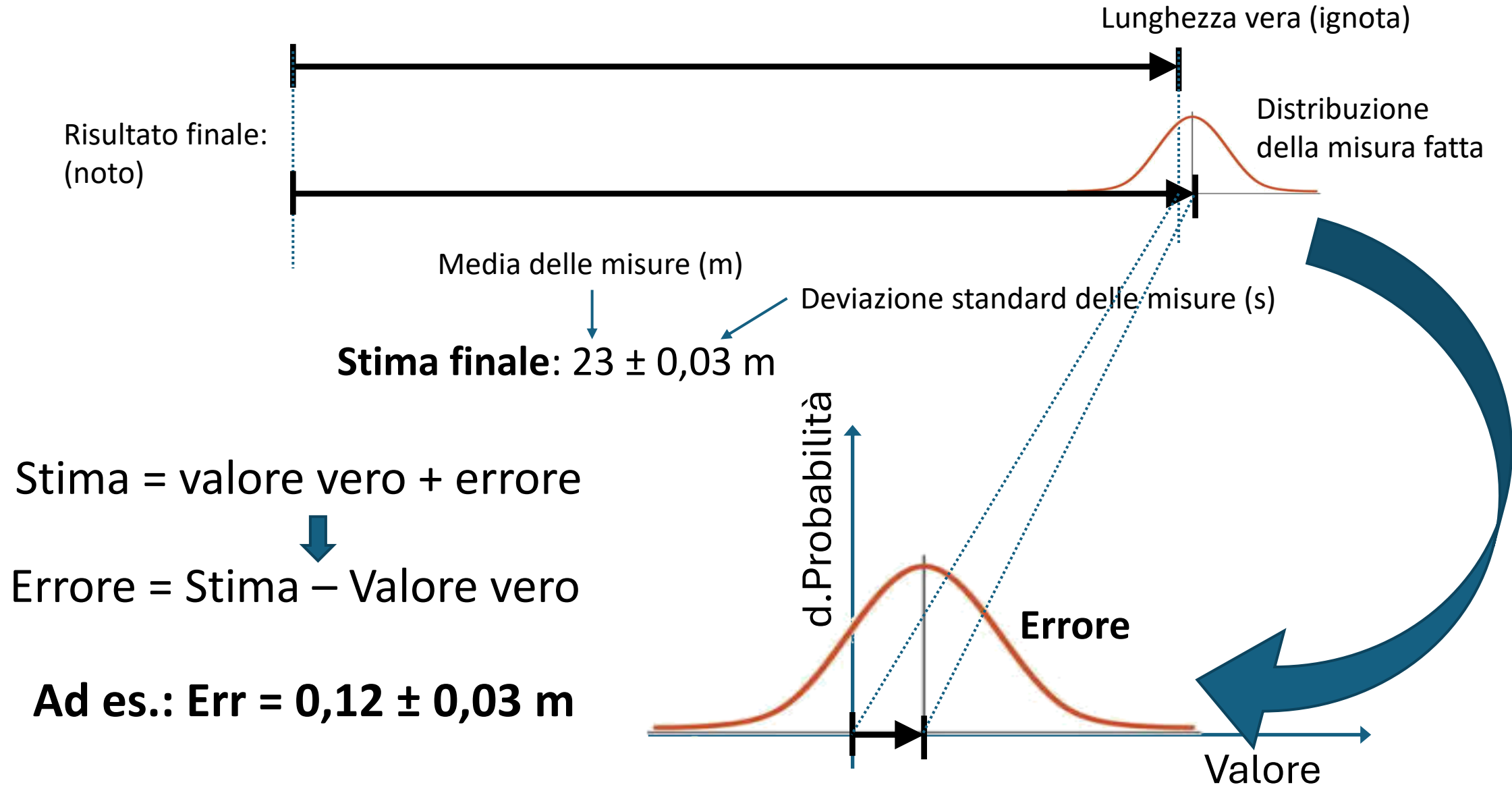
Non possiamo conoscere il valore dell'errore ma possiamo ottenere le sue **caratteristiche statistiche**, ad es. media e deviazione standard.

L'ERRORE

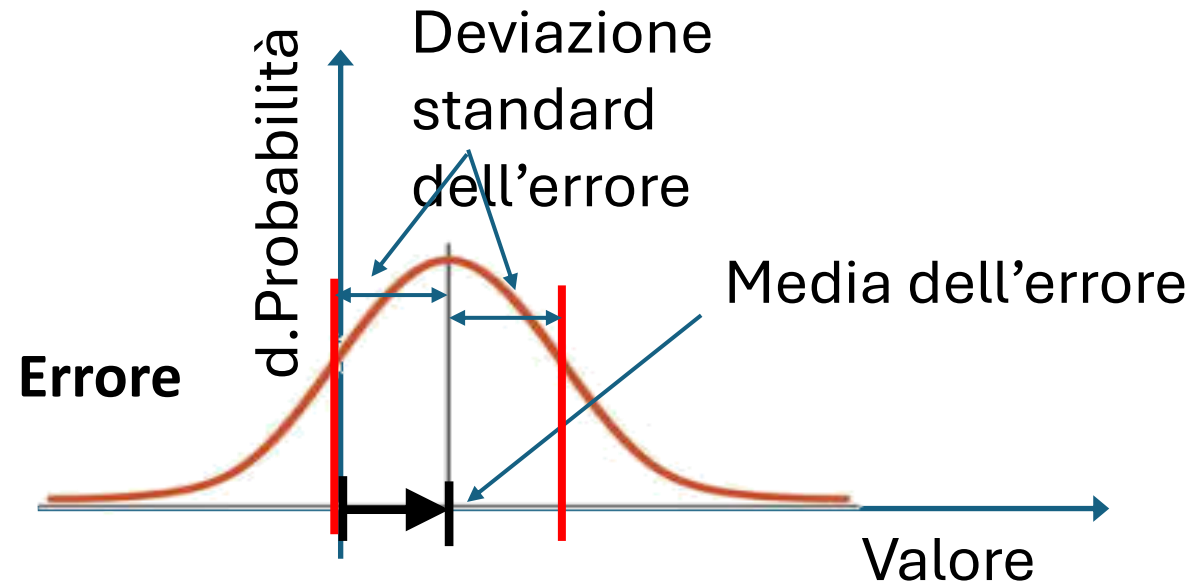
Esempio con le lunghezze ma vale con tutto



L'ERRORE



L'ERRORE



Nell'esempio il valore ignoto dell'errore è:

$$\text{Err} = 0,12 \pm 0,03 \text{ m}$$

Media dell'errore
(**ignota**)

Deviazione standard
dell'errore, uguale alla
DS delle misure
(**spesso misurabile**)

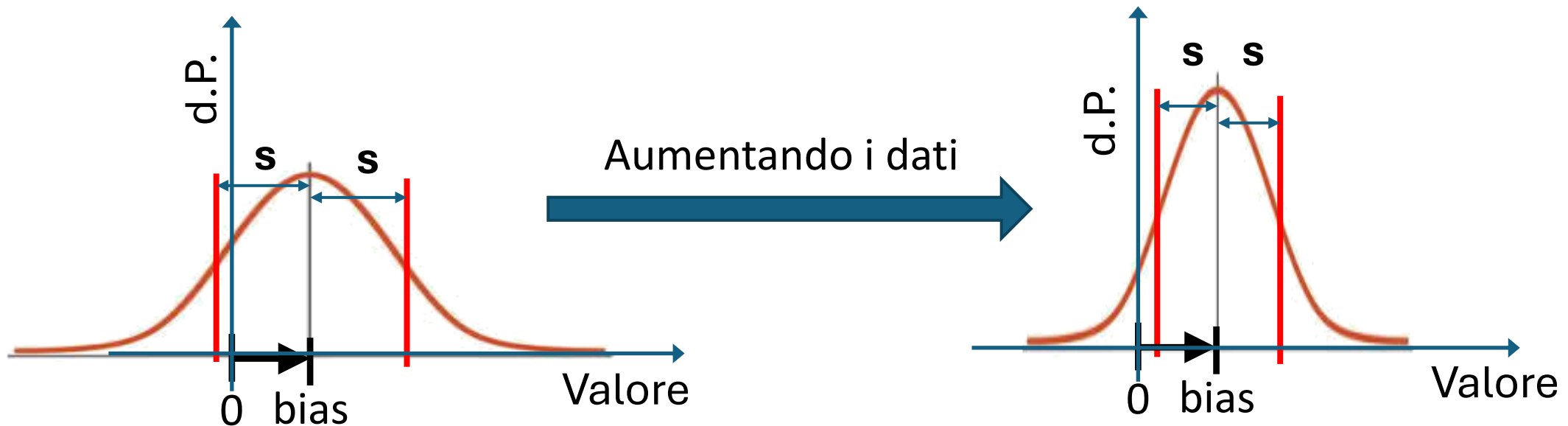
La media dell'errore prende il nome di **bias**
(pronunciato 'baias')

L'ERRORE – la deviazione standard

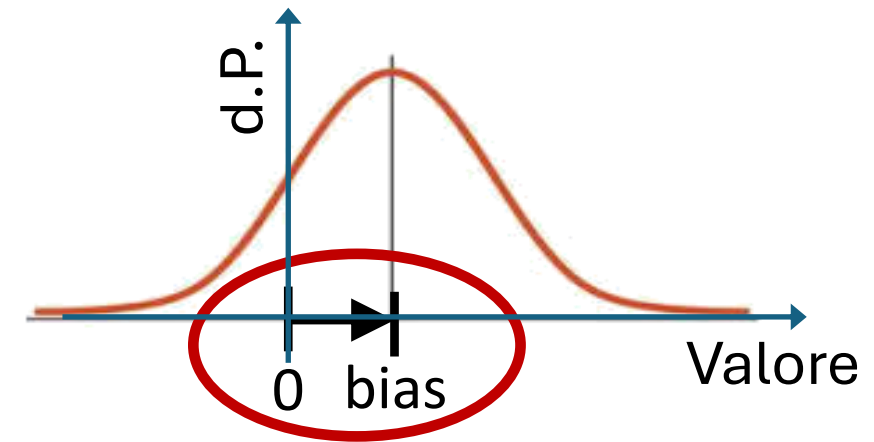
Deviazione standard dell'errore

Aumentare il numero delle misure, o dei dati, riduce l'incertezza della stima, cioè la deviazione standard dell'errore finale (vedi «errore standard della media»).

Lo stesso vale per l'aumento dei soggetti testati nei trial clinici o negli studi osservazionali, ecc.



L'ERRORE - bias



Non c'è modo semplice per conoscere o ridurre il **bias**

Si opta quindi per tecniche di misura o formule di stima appositamente studiati per ridurre il più possibile il bias, sempre (a meno che non si voglia falsificare i risultati). Accortezze per azzerare il bias vengono sistematicamente applicate e si può notare in molte pratiche comuni.

Ad esempio:

- Quando il medico ti fa pesare ti fa spogliare per evitare di contare il peso dei vestiti che andrebbero ad aumentare la misura, indipendentemente dalla precisione della bilancia.
- Se misuro una distanza col metro flessibile, faccio in modo che il metro sia teso o la misura aumenta indipendentemente dalla precisione del metro.

L'ERRORE - bias

Come misuriamo la lunghezza di un tavolo?

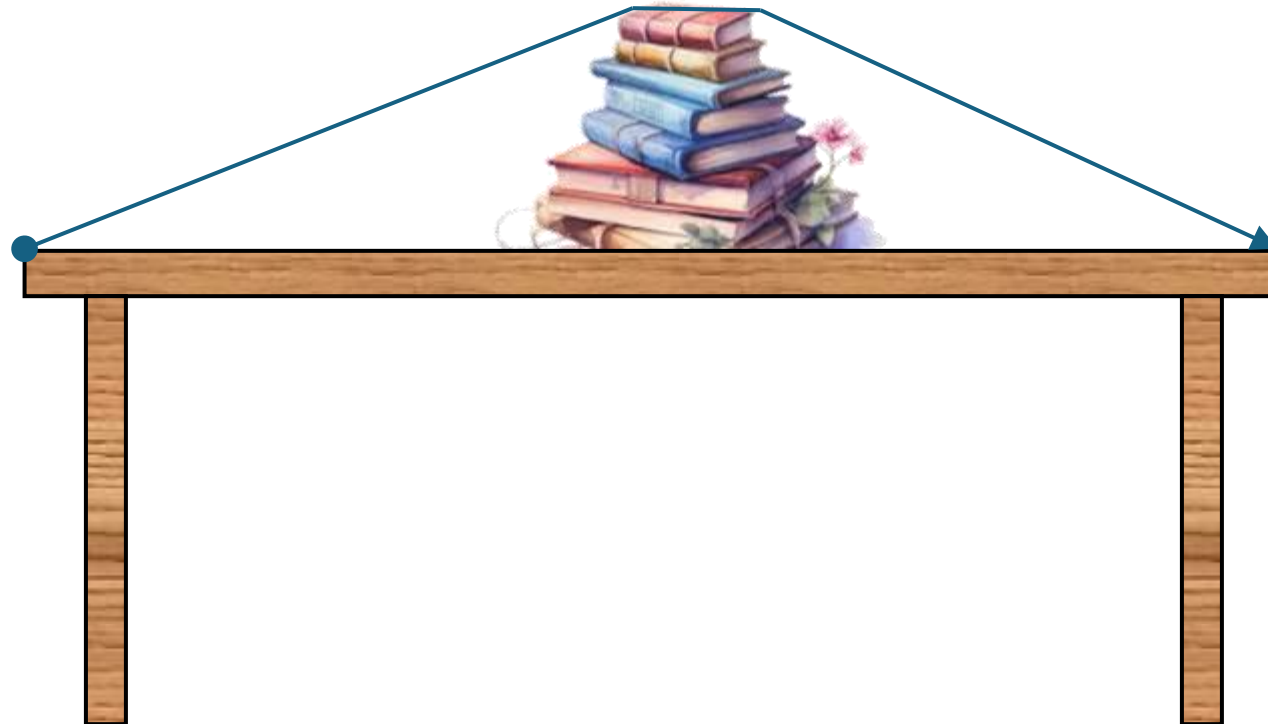


L'ERRORE - bias

Il nuovo ministro del commercio, proprietario di una falegnameria, dice che la misura per la lunghezza dei tavoli negli uffici pubblici va presa con 10 copie della costituzione, del codice civile e di quello penale, impilate nel mezzo del tavolo.

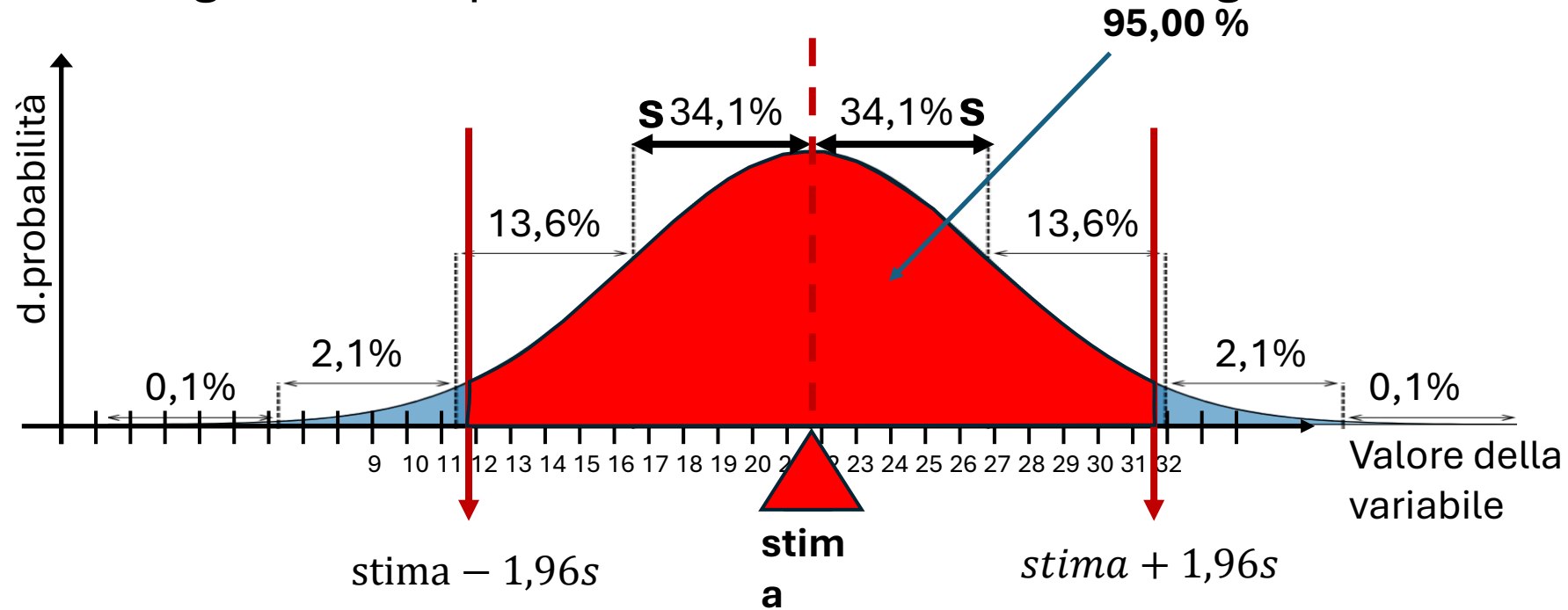
Il metro va fatto passare sopra ai libri in omaggio allo Stato. È il protocollo.

Chi non lo fa è un nemico della patria e non puoi lavorare per lo Stato!



Intervallo di confidenza

Se ho una stima, o una misura, e conosco l'incertezza (deviazione standard dell'errore, s) (**e il bias è nullo o trascurabile**) e il tipo della distribuzione di probabilità dell'errore (ad es la normale) posso definire un intervallo attorno alla stima che contenga con alta probabilità il valore vero che voglio stimare.



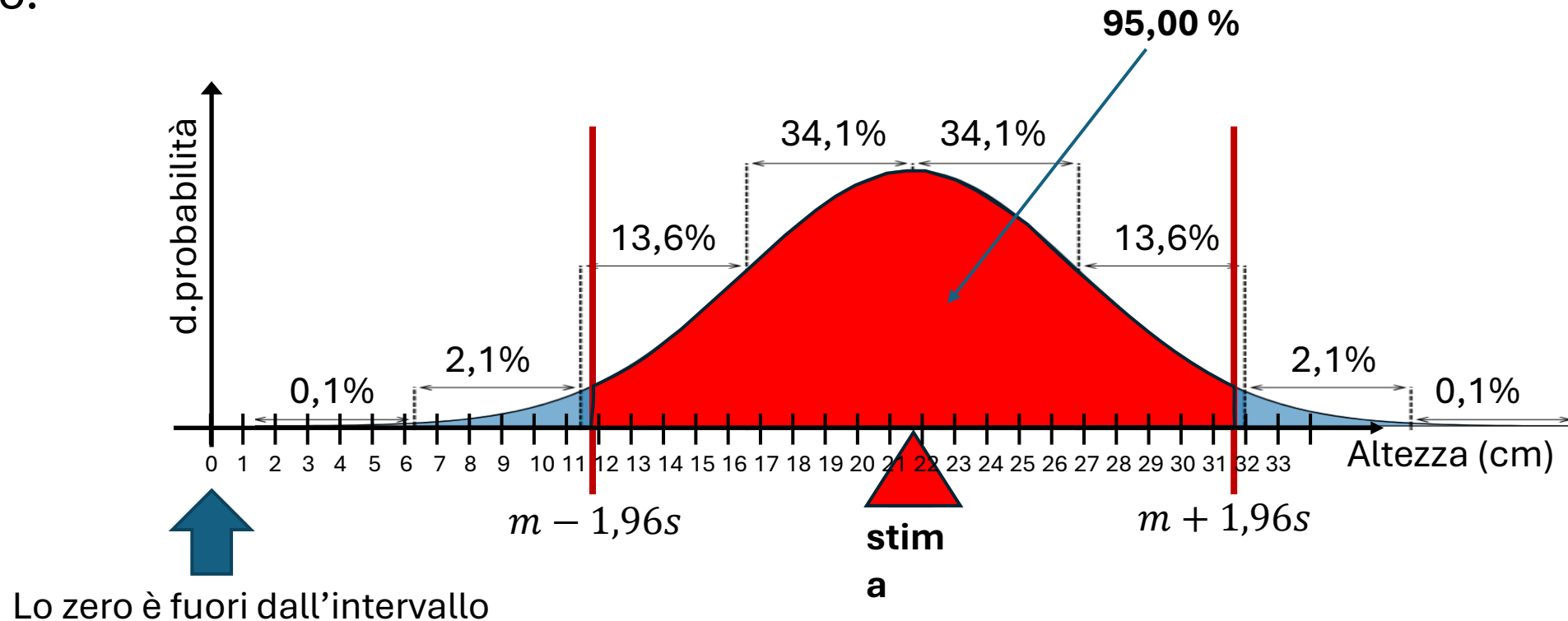
Nell'esempio si usa una normale, e nell'intervallo da stima -1,96xS a stima +1,96xS

ho il 95% di probabilità che ci sia il **valore vero** che la stima voleva identificare.

Intervallo di confidenza - significatività

Una stima ha un significato solo se è sufficientemente precisa.

Il livello minimo di precisione è quello di sapere se il valore vero è positivo o negativo.

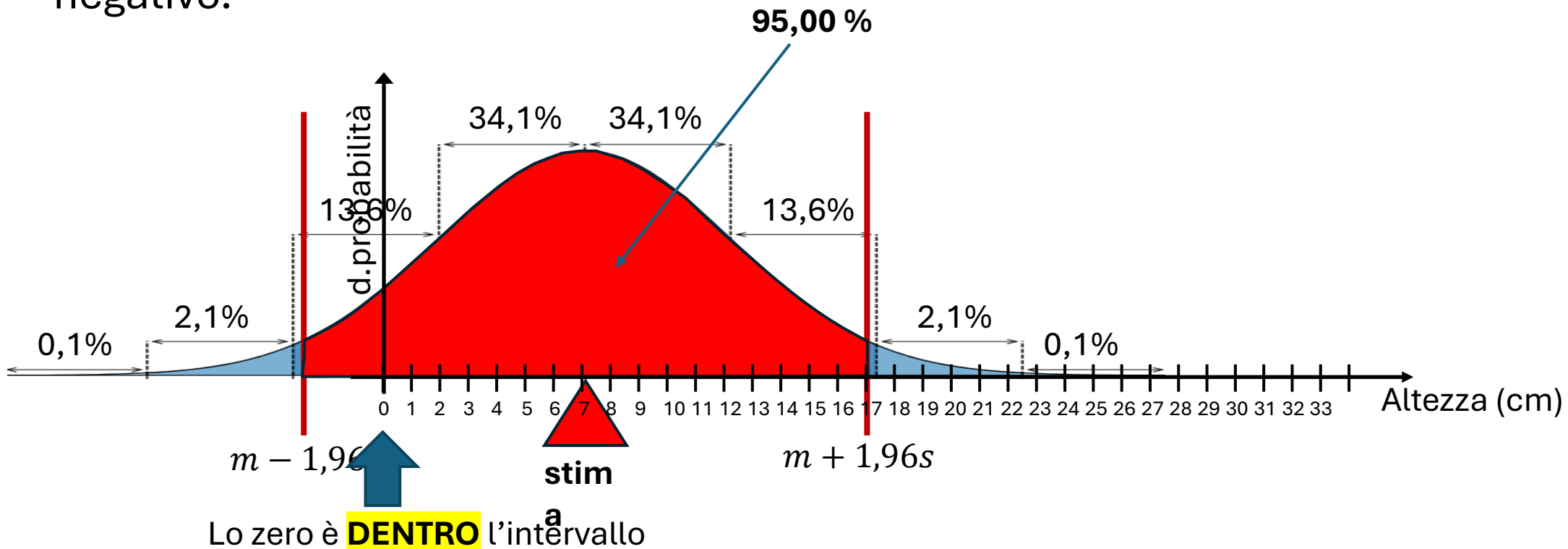


In questo caso la stima è significativa perché l'intervallo di confidenza non comprende lo zero, è tutto positivo. Andrebbe bene anche se fosse tutto negativo. ²⁸

Intervallo di confidenza - significatività

Una stima ha un significato solo se è sufficientemente precisa.

Il livello minimo di precisione è quello di sapere se il valore vero è positivo o negativo.



Col 95% di certezza non so neppure se il valore vero è positivo o negativo. La stima non è significativa. (Nella differenza da 1 dei rapporti va usato l'uno al posto dello zero)

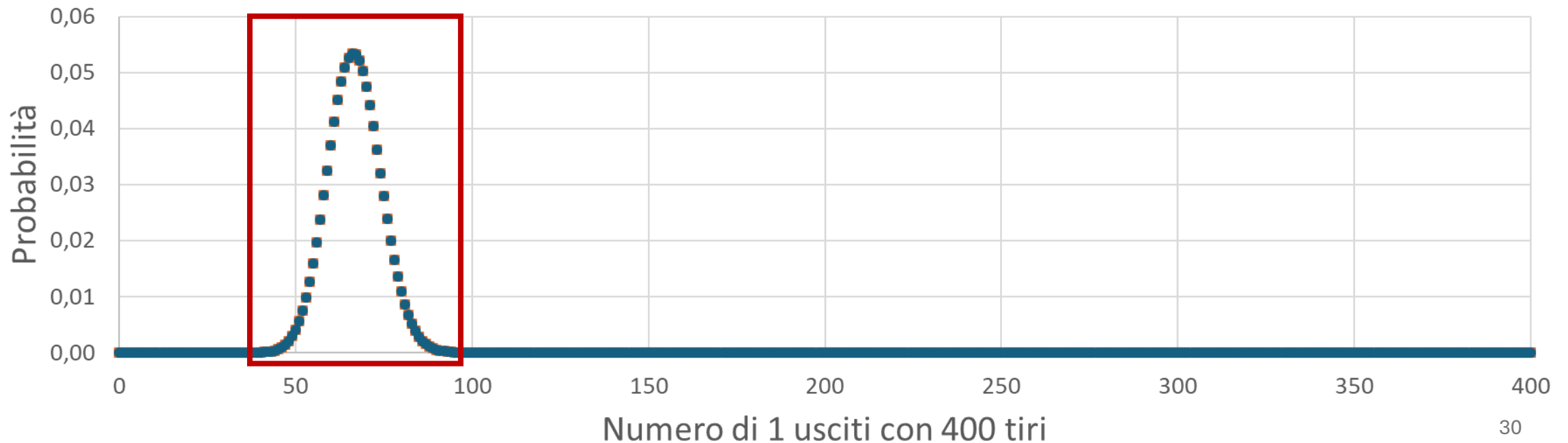
Test statistico – esempio coi miei dadi

Ma è mai possibile fare 86 di uno con 400 tiri per caso se il dado non è truccato?

Vado a vedere quale distribuzione di probabilità posso usare per i tiri di dadi, **n** tiri, **k** valori uguali usciti, **p** probabilità di uscire ($1/6$).

È la distribuzione binomiale (assomiglia alla normale ma è per numeri interi). Con 400 tiri, la probabilità che esca 1 un dato numero di volte, con probabilità $1/6$, è quella qui sotto.

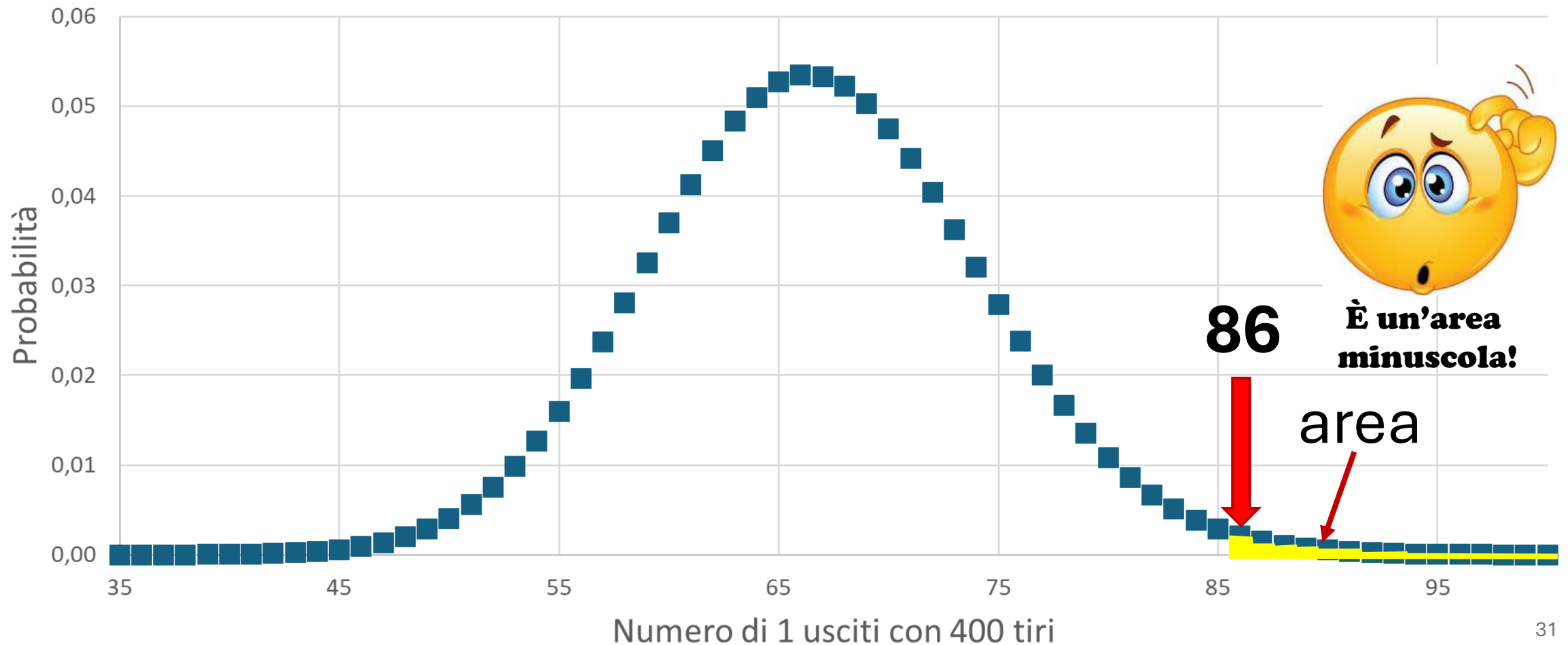
Distribuzione binomiale (per numeri interi), $n = 400$, $k = 1, \dots, 400$



Test statistico

Ingrandimento nella zona di interesse

Distribuzione binomiale (per numeri discreti) , $n = 400$, $k = 1, \dots, 400$



Test statistico - il p-value

Domanda statistica corretta:

Quale è la probabilità che, tirando 400 volte dei dadi bilanciati (per cui si fa 1 una volta su sei) vengo ad avere 86 volte uno o di più?

Risposta statistica:

- sotto l'ipotesi (**H0**) usata per fare i conti (cioè che il dado sia bilanciato),
- la probabilità di fare 86 o più volte uno con 400 tiri (area della binomiale sopra a 85) è:
- (con Excel formula: $=1 - \text{DISTRIB.BINOM.N}(86-1; 400; 1/6; 1)$) = **0,7%**

Risposta pratica:

Abbiamo lo 0,7% di possibilità che i miei dadi siano bilanciati e che sia solo un caso che sembra facciano tante volte 1... **Quindi i miei dadi hanno il 99,3% di non essere bilanciati.**

P-value

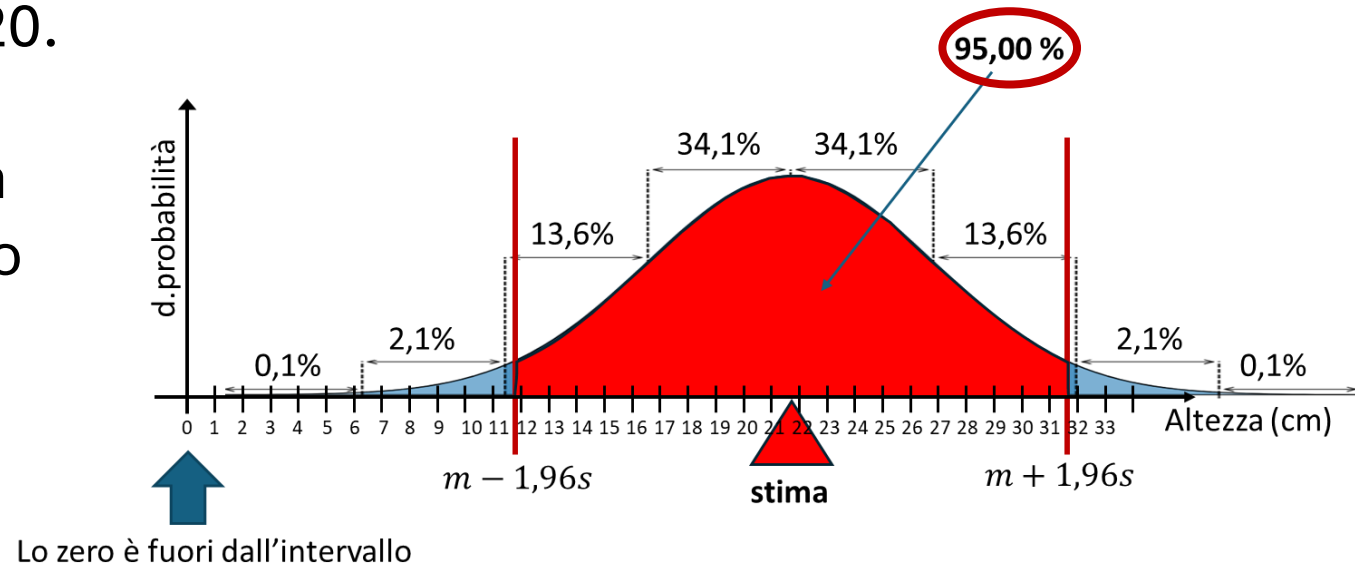
È la probabilità di ottenere il caso testato se **H0** è vera. Mi serve per verificare quanto l'ipotesi regge alla prova dei fatti.

Test statistico - il p-value

Quando il p-value è abbastanza basso da considerare falso H_0 ?
(questo valore viene chiamato «**alfa**»)

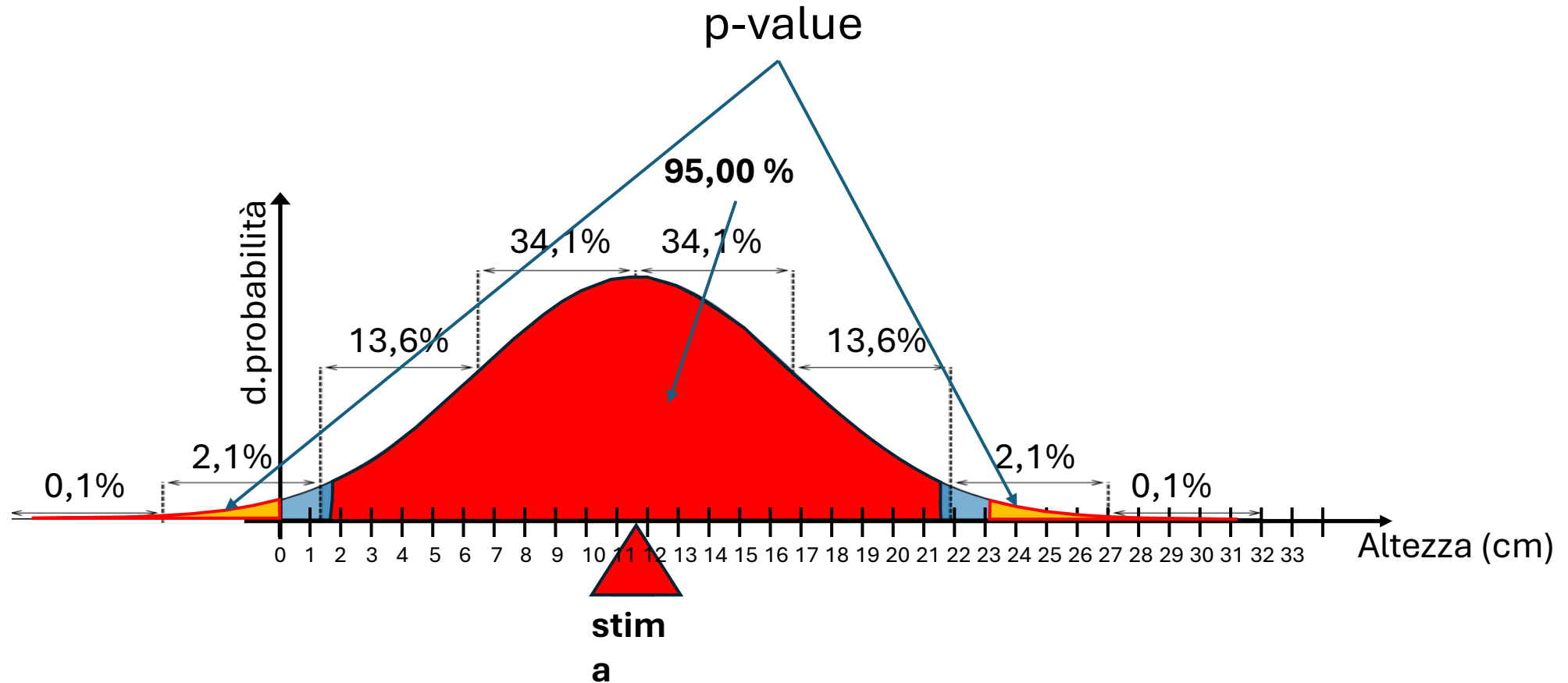
SE ci sono **fondati sospetti** che H_0 sia falso (ad es spiegazione scientifica del fenomeno) e il test è stato **fatto apposta per dimostrarlo** allora basta sia minore del **5%**.

Posso sbagliarmi per caso 1 volta su 20.
È per questo che si usa spesso
l'intervallo di confidenza al **95%** per la
significatività, l'avanzo al 100%, quello
fuori dall'area rossa, impone che il **p-value**
sia inferiore al 5%.



Test statistico - il p-value

In questo caso (test a doppia coda) il p-value è l'area che sta sotto allo zero più l'area simmetrica sull'altra coda della curva.



Se lo zero non è nella zona rossa (95%) allora, per forza, area arancione è meno del 5% dell'area totale della curva.

Riassunto

1. La probabilità è quanto sia facile che una cosa capiti (0 – 100%)
2. La media è il punto di mezzo di tutti i dati raccolti
3. La deviazione standard è una misura di variabilità dei valori attorno alla media
4. L'istogramma di frequenza ci dice quanto è frequente che dei valori siano osservati
5. Se ho abbastanza misure l'istogramma assomiglia alla distribuzione di probabilità
6. La distribuzione di probabilità più comune è la NORMALE
7. Con la normale si può calcolare la probabilità (area) che un valore ricada entro un certo intervallo dalla media.
8. Ci sono grosse differenze tra misura diretta, misura indiretta e stima. Occorre che tutte le **condizioni** siano verificate o i risultati possono essere distorti

Riassunto

- 10. Il «**BIAS**» è la media dell'errore e devono essere prese precauzioni perché sia zero oppure sono guai seri perché del bias non me ne libero.
- 11. Usare protocolli di misura che producano bias è un ottimo modo per manipolare i dati
- 12. L'**intervallo di confidenza** e la «**significatività statistica**»
- 13. Un esempio di test statistici per la verifica delle ipotesi e il «**p-value**»

Un esempio di bias nella stima dell'efficacia vaccinale

Da: "L'analisi dell'efficacia del vaccino anti-SARS-CoV-2 con studi sulla popolazione generale. Gli errori metodologici che provocano un aumento apparente dell'efficacia" del Prof. Stefano Petti - Professore Ordinario, Dipartimento di Sanità Pubblica e Malattie Infettive, Università "La Sapienza", Roma

Stima dell'efficacia vaccinale con il vaccino covid19 ottenuta usando i dati «numero di contagiati» e «numero di vaccinati»

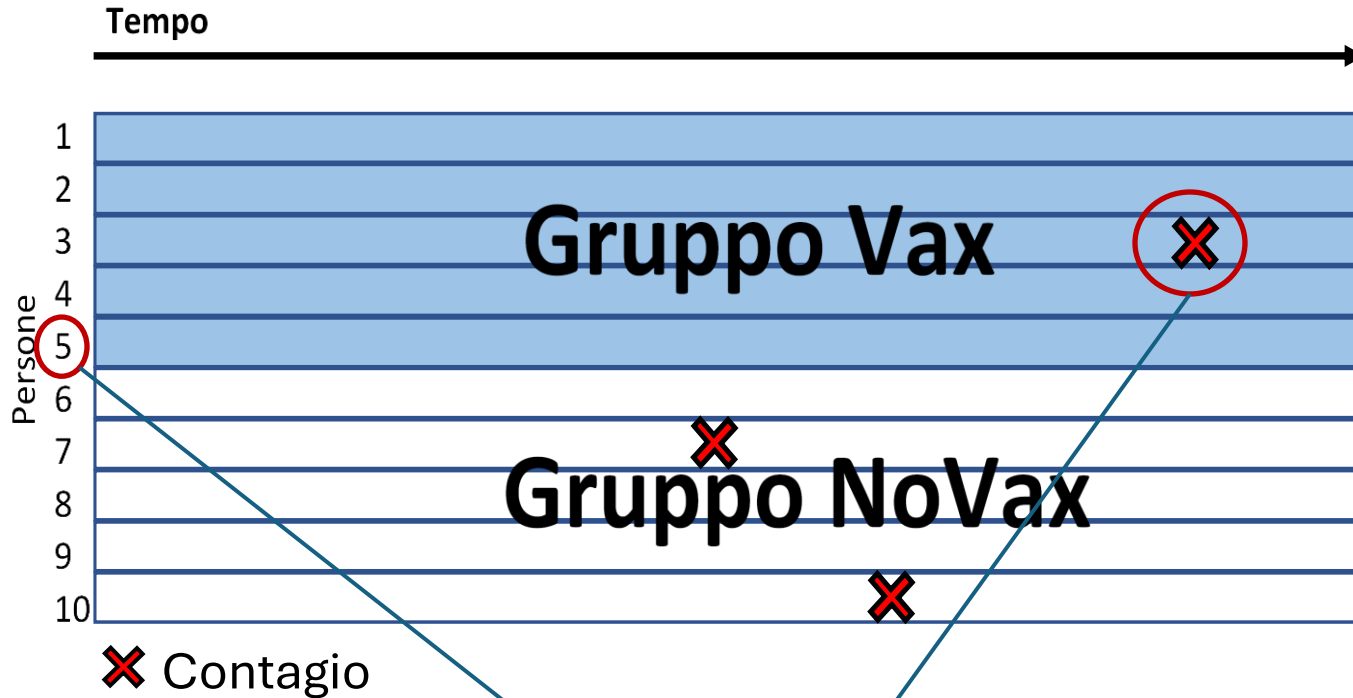
In mancanza dei dati necessari per applicare l'approccio per lo «studio osservazionale» hanno utilizzato l'approccio per il «**trial clinico**» a cui bastano:

- numero di vaccinati
- numero di contagiati tra i vaccinati e tra i non vaccinati

Vediamo nell'ordine:

- 1) Come è fatto l'approccio del trial clinico
- 2) Come è stato **erroneamente** applicato quello stesso approccio allo studio osservazionale covid19
- 3) Come sono i risultati del medesimo esempio se si usa l'approccio corretto

Trial clinico (corretto)



Incidenza

$$i.V = \frac{\text{Casi}}{\text{Numero di persone}} \times 100 = \frac{1}{5} \times 100 = 20\%$$

$$i.NV = \frac{\text{Casi}}{\text{Numero di persone}} \times 100 = \frac{2}{5} \times 100 = 40\%$$

Incidenza

che percentuale di gente si è ammalata?

$$i = \frac{\text{Casi}}{\text{Numero di persone}} \times 100$$

Efficacia vaccinale

quale riduzione % della incidenza dato dal vax?

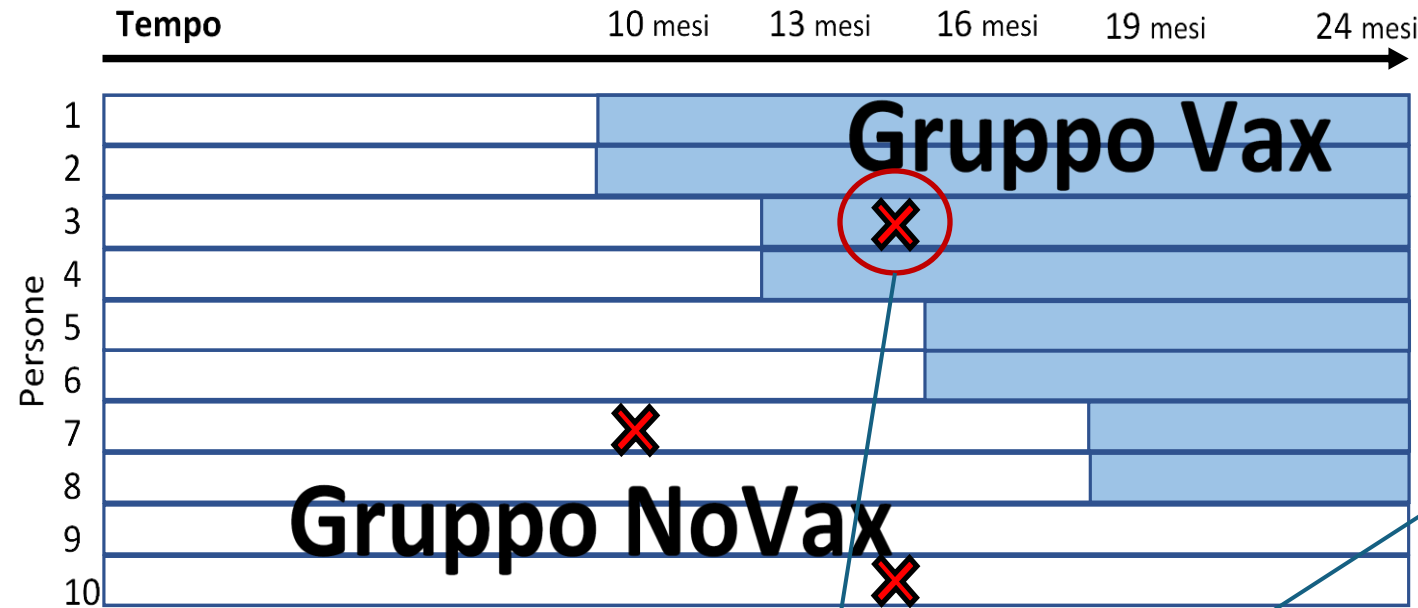
$$EV = \frac{i.NV - i.V}{i.NV} \times 100$$

Efficacia vaccinale

$$EV = \frac{i.NV - i.V}{i.NV} \times 100 = \frac{40 - 20}{40} \times 100 = 50\%$$

In questo caso il vaccino dimezza i casi

Studio osservazionale covid19 con approccio errato



Incidenza (trial clinico)

che percentuale di gente si è ammalata?

$$i = \frac{\text{Casi}}{\text{Numero di persone}} \times 100$$

Efficacia vaccinale (trial clinico)

quale riduzione % della incidenza dato dal vax?

$$EV = \frac{i.NV - i.V}{i.NV} \times 100$$

Incidenza (per trial clinico)

$$i.V = \frac{\text{Casi}}{\text{Numero di persone}} \times 100 = \frac{1}{8} \times 100 = 12,5\%$$

$$i.NV = \frac{\text{Casi}}{\text{Numero di persone}} \times 100 = \frac{2}{2} \times 100 = 100\%$$

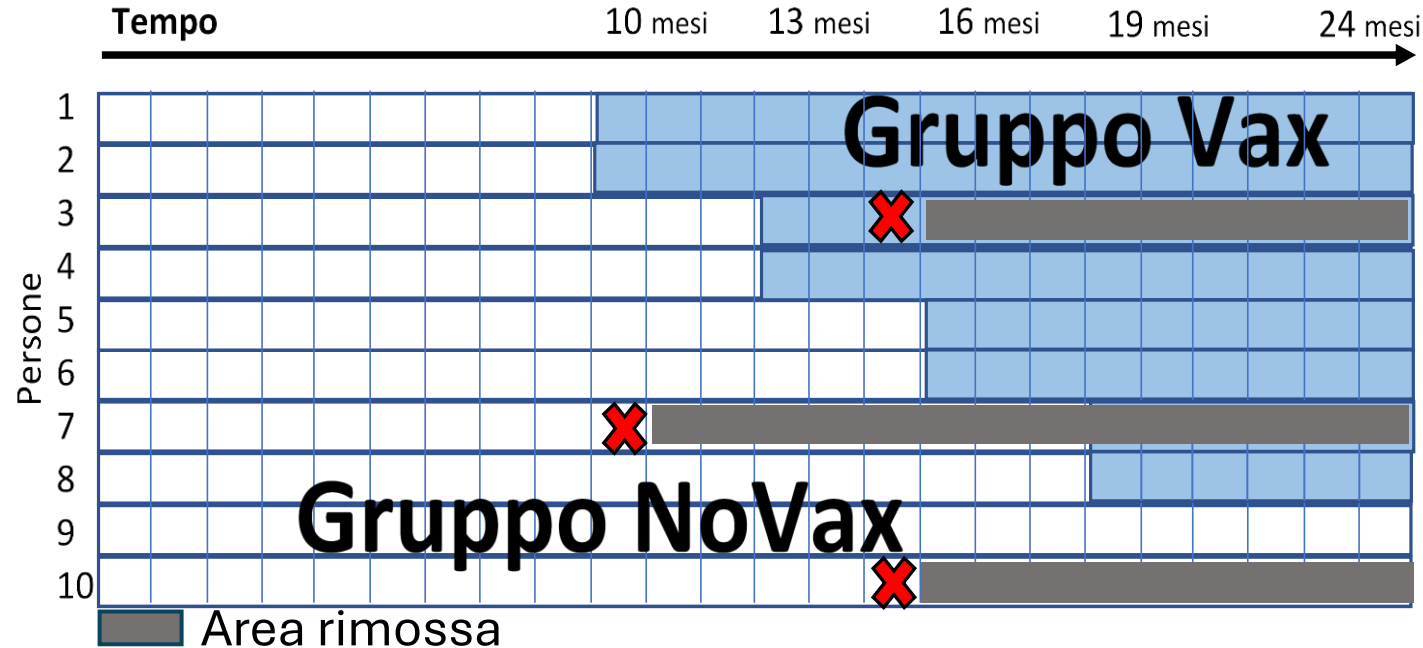
Efficacia vaccinale (per trial clinico)

$$EV = \frac{i.NV - i.V}{i.NV} \times 100 = \frac{100 - 12,5}{100} \times 100 = 87,5\%$$

riduce i casi del **90%**.

Bias dovuto a tempi di esposizione differenti!

Studio osservazionale corretto



Aree «tempo-persona» (quadretti in figura)

Area Vax = 15 + 15 + 3 + 12 + 9 + 9 + 6 = 69 mesi-persona

Area NoVax = 9 + 9 + 12 + 12 + 15 + 15 + 10 + 18 + 24 + 15 = 139 mesi-persona

Casi

Casi Vax = 1

Casi NoVax = 2

Densità di incidenza

$$di.V = \frac{casi}{area} \times 100 = \frac{1}{69} \times 100 = 1,45\%$$

$$di.NV = \frac{casi}{area} \times 100 = \frac{2}{139} \times 100 = 1,43\%$$

Calcolo efficacia vaccinale esatto (con densità di incidenza per studi osservazionali)

$$EV = \frac{di.NV - di.V}{di.NV} \times 100 = \frac{1,43 - 1,45}{1,43} \times 100 = -1,4\%$$

aumenta i casi del **1%**

Misura diretta, indiretta e stima

Le sorgenti di incertezza nelle misure

Stima - è una misura indiretta ottenuta attraverso qualche tipo di approccio statistico.

La stima è basata su solitamente su un «modello statistico» e il valore risultante è soggetto ad errore statistico che dipende da:

- errori accidentali informatici o di battitura
- errori nelle informazioni (dati in ingresso)
- semplificazioni del modello

← **errori grossolani**

← **incertezze** (simili alle misure indirette)

Inoltre il modello adottato è valido solo se:

- i dati rispettano le condizioni di validità del modello
- il caso esaminato combacia esattamente con quello ipotizzato nello sviluppo delle equazioni del modello.

← **condizioni di validità** (simili alle misure indirette)

← **ambito di applicazione**

Da 87% a -1%

PRO: si possono ottenere stime di cose che altrimenti è impossibile misurare direttamente o indirettamente.

CONTRO: occorre prestare sempre molta attenzione a che tutte le condizioni siano valide o, anche in buona fede, vengono fuori valori totalmente fuorvianti con poche possibilità di riscontro.

Grazie per l'attenzione!

