Statistica di base

- ***un problema di calcolo delle probabilità*: se abbiamo $p = 1/2$ di ottenere testa in un lancio, qual è la probabilità di ottenere in 1000 lanci meno di 450 teste?**
- ***lo stesso problema in statistica*: se in 1000 lanci si ottengono 450 teste, qual è la stima della probabilità vera di ottenere testa?**

Errore statistico: $s \approx \sigma$

$$\mu \pm \sigma = 500.0 \pm 15.8 \simeq 500 \pm 16 = [484, 516]$$
$$x \pm s = 450.0 \pm 15.7 \simeq 450 \pm 16 = [434, 466]$$

CALCOLO DELLE PROBABILITÀ	STATISTICA
probabilità di valori dello spettro	**stima di parametri** ($\hat{p}$)
probabilità vera: $p = 0.5$	**frequenza:** $f = x/n = 0.45$
valore atteso: $\langle X \rangle =$ **500**	**valore misurato:** $x = 450$
deviazione standard: $\sigma[X] = \sqrt{np(1-p)} = 15.8$	**errore statistico o incertezza:** $s = \sqrt{nf(1-f)} = 15.7$

Statistica di base

La statistica comprende due tipi di inferenza:

- **stima di parametri: stimare la probabilità da 1000 lanci**
- **verifica di ipotesi: se ripetendo due volte l'esperimento dei mille lanci si ottengono 450 e 600 successi, quanto è probabile che nei due esperimenti si sia usata la stessa moneta?**

Le leggi in statistica in genere dipendono da set di parametri θ:

$$\mathcal{E}(\theta) \equiv (S, \mathcal{F}, P_\theta)$$

corrispondente ad una densità:

$$P\{X \in A\} = \int_A p(x;\theta)\,\mathrm{d}x$$

In fisica sperimentale

- **la massa dell'Higgs vale (PDG 2000):**

$$m > 95.3\ GeV, CL = 95\%$$

- **massa del W:**

$$m_W = 80.419 \pm 0.056\ GeV$$

Questi sono gli
intervalli di confidenza

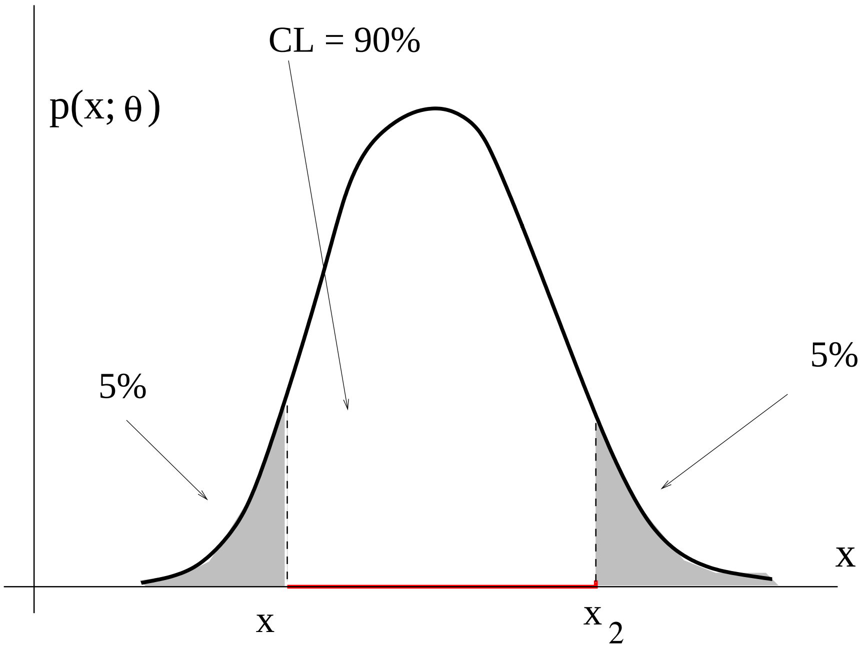
COSA SIGNIFICANO?

Intervalli di confidenza

Si parte dal calcolo delle probabilità

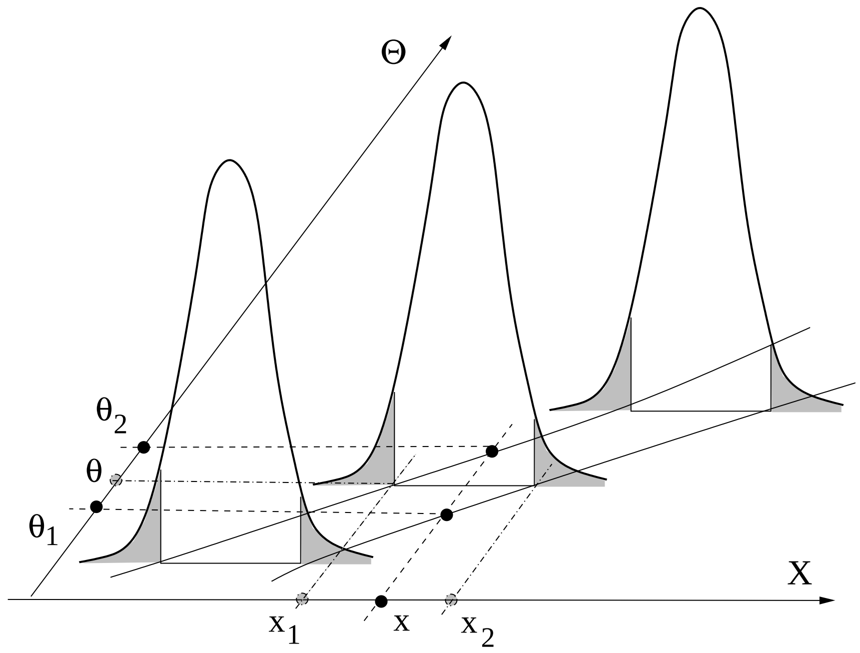
$$\int_{x_1}^{x_2} p(x;\theta)\,\mathrm{d}x = CL$$

Si ripete il procedimento per



tutti i possibili valori di θ

Intervalli di confidenza



Dalla figura risulta:

$$X \in [x_1, x_2] \text{ se e solo se } \Theta \in [\theta_1, \theta_2]$$

Poichè

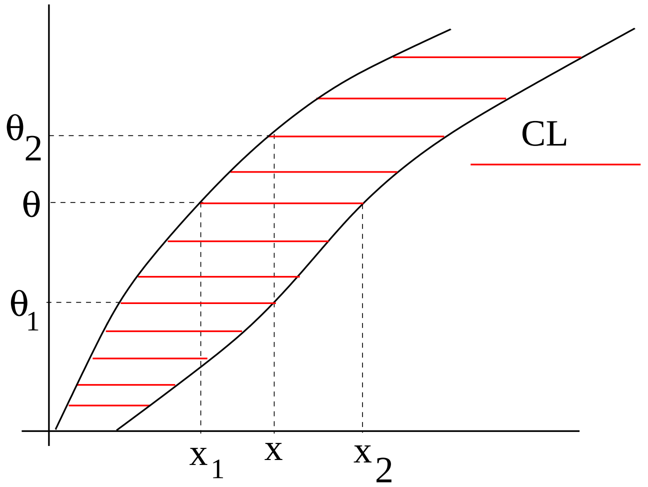
$$P\{X \in [x_1, x_2]\} = CL$$

allora

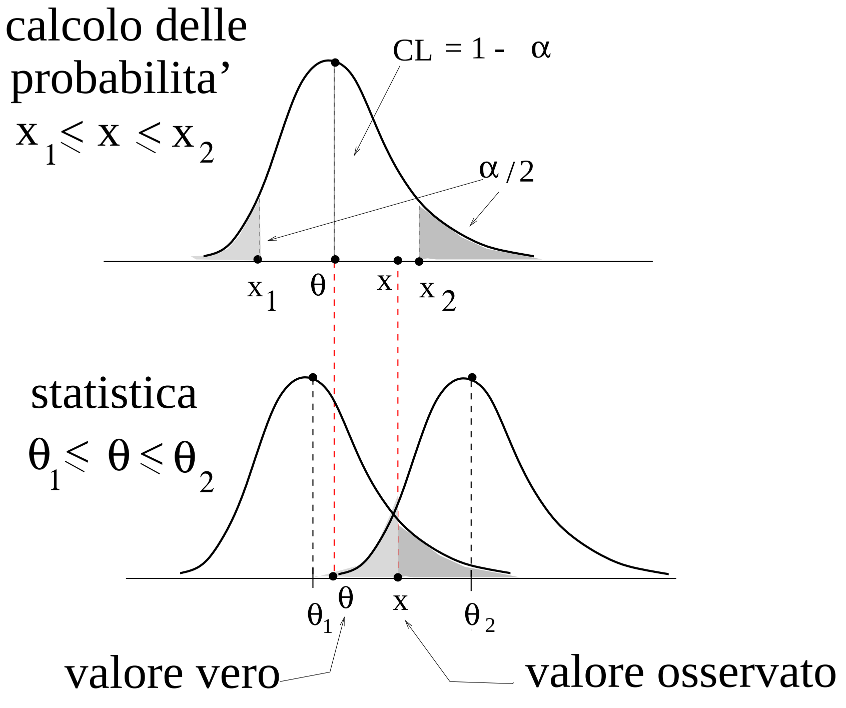
$$P\{\Theta \in [\theta_1, \theta_2]\} = CL$$

Fondamentale risultato di Neyman (1937)

Neyman dall'alto e di fronte:



$$P\{\theta_1 < \theta < \theta_2\} = P\{x_1 < x < x_2\} = CL$$



Intervalli di confidenza definizione matematica

Date due statistiche T_1 e T_2 si dice che

$$I = [T_1, T_2]$$

è un intervallo di confidenza o di fiducia per un parametro θ, di livello di confidenza $0 < CL < 1$, se, per ogni $\theta \in \Theta$ la probabilità che I contenga θ (*coverage*) vale CL:

$$P\{T_1 \leq \theta \leq T_2\} = CL \ .$$

Se T_1 e T_2 sono variabili discrete, l'intervallo di confidenza è soddisfa il *minimum overcoverage*

$$P\{T_1 \leq \theta \leq T_2\} \geq CL \ .$$

$[T_1, T_2]$ sono variabili aleatorie, mentre il parametro θ è fissato

Intervalli di confidenza
definizioni equivalenti

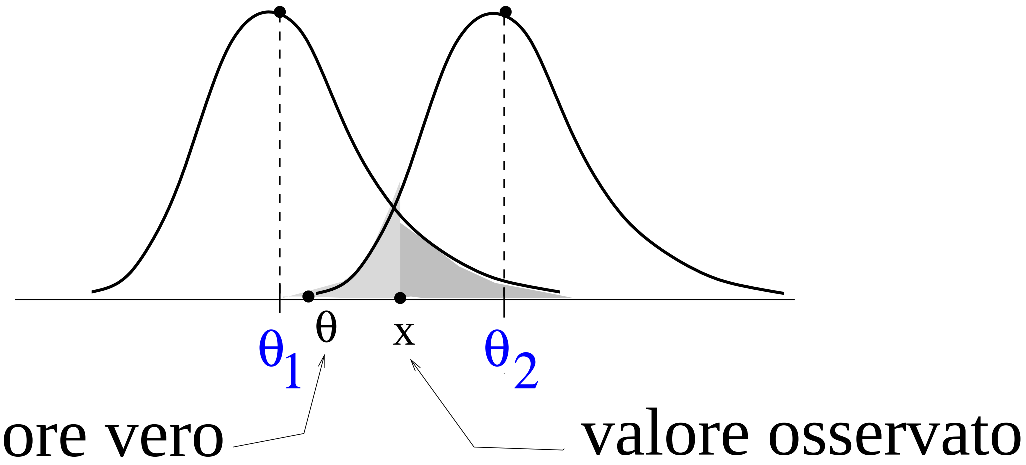
Definizioni giuste:

- CL è la probabilità che l'intervallo aleatorio $[T_1, T_2]$ comprenda il valore vero;
- in un insieme infinito di esperimenti identici ripetuti, una frazione CL di questi avrà successo nel localizzare il valore vero come $\theta \in [\theta_1, \theta_2]$;
- se $\theta \notin [\theta_1, \theta_2]$, posso ottenere $\{I = [\theta_1, \theta_2]\}$ ma in una frazione di esperimenti $\leq 1 - CL$
  Questa definizione è quella valida per i limiti superiori e inferiori

Definizioni sbagliate:

- CL è la probabilità che il valore vero sia in $[\theta_1, \theta_2]$.
- CL è la probabilità che $\theta \in [\theta_1, \theta_2]$

Determinazione degli intervalli



$$\int_x^\infty p(x;\theta_1)\,\mathrm{d}x = c_1 \qquad \int_{-\infty}^x p(x;\theta_2)\,\mathrm{d}x = c_2$$

dove

$$\theta \in [\theta_1, \theta_2]\ , \quad 1-(c_1+c_2) = CL$$

Si usano anche tecniche MC: griglia di valori di θ per trovare i valori θ_1 e θ_2 che soddisfano gli integrali

importante:

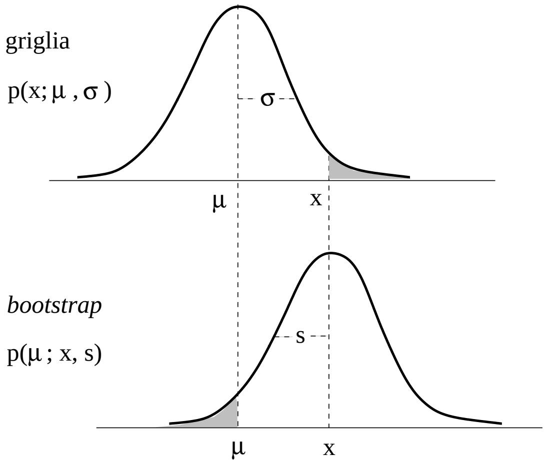
$$\int_{\theta_1}^{\theta_2} p(\theta;x)\,\mathrm{d}\theta = CL$$

ERRATO!!!!

Determinazione degli intervalli

Quando vale la proprietà

$$1 - \int_{-\infty}^{x} p(x;\theta)\,\mathrm{d}x = \int_{-\infty}^{\theta} p(\theta;x)\,\mathrm{d}\theta$$



$$\int_{x}^{\infty} p(x;\theta_1)\,\mathrm{d}x = c_1 \quad \int_{-\infty}^{x} p(x;\theta_2)\,\mathrm{d}x = c_2$$

$$c_1 = \int_{x}^{\infty} \ldots = 1 - \int_{\theta_1}^{\infty} \ldots = \int_{-\infty}^{\theta_1} p(\theta;x)\,\mathrm{d}\theta$$

$$c_2 = \int_{-\infty}^{x} \ldots = 1 - \int_{-\infty}^{\theta_2} \ldots = \int_{\theta_2}^{\infty} p(\theta;x)\,\mathrm{d}\theta$$

si può integrare in θ in modo bayesiano. (Fonte di confusione!)

$$CL = 1 - c_1 - c_2 = \int_{\theta_1}^{\theta_2} p(\theta;x)\,\mathrm{d}\theta$$

Determinazione degli intervalli

La proprietà precedente vale solo se

$$1 - F(x;\theta) = F(\theta;x)$$

cioè per densità simmetriche in θ invarianti per traslazione, come la gaussiana:

$$p(x;\theta) \propto \exp\left[-\frac{1}{2}\frac{(x-\theta)^2}{\sigma^2}\right]$$

da cui

$$P\{\mu-\sigma \le X \le \mu+\sigma\} = P\{-\sigma \le X-\mu \le \sigma\}$$
$$= P\{X-\sigma \le \mu \le X+\sigma\}$$

Un integrale come

$$\int_{\theta_1}^{\theta_2} p(\theta;x)\,\mathrm{d}\theta$$

è la stima di Bayes con probabilità a priori uniforme.

Quantità pivot

Metodo che evita di risolvere gli integrali $\int_A p(x;\theta)\,\mathrm{d}x = c_i$

Se $Q(x;\theta)$ è pivotale, $P\{Q \in A\}$ non dipende da θ. Esempio:

$$Q = (X - \theta) \sim N(0, \sigma^2)$$

Metodo:

- **trovare $P\{q_1 \le Q \le q_2\} = CL$;**
- **invertire la relazione:**

$$Q(x;\theta) = q \rightarrow \theta = T(x;q)$$

- **Allora:**

$$P\{q_1 \le Q \le q_2\} = P\{T_1 \le \theta \le T_2\} = CL$$

Metodo comodo....

ma fa scordare il vero significato delle formule!

Stima della probabilità piccoli campioni

... qui cominciano i dolori ....
mancano quantità pivot:

$$\sum_{k=x}^{n} \binom{n}{k} p_1^k (1-p_1)^{n-k} = c_1 \; ,$$

$$\sum_{k=0}^{x} \binom{n}{k} p_2^k (1-p_2)^{n-k} = c_2 \; .$$

Caso simmetrico: $c_1 = c_2 = (1-CL)/2 = \alpha/2$.

Quando $x = 0$, $x = n$, $c_1 = c_2 = 1 - CL$:

$$\begin{aligned} x = n \implies \quad & p_1^n = 1 - CL \; , \\ x = 0 \implies \quad & (1-p_2)^n = 1 - CL \; . \end{aligned}$$

quando tutti i tentativi hanno avuto successo:

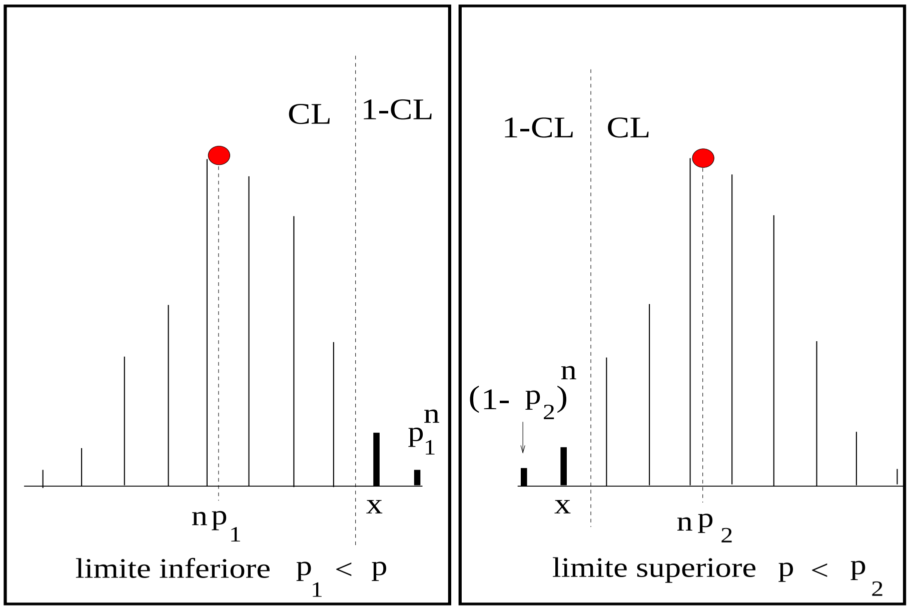
$$p_1 = \sqrt[n]{1-CL} \qquad p \in [p_1, 1]$$

quando non si è registrato alcun successo:

$$p_2 = 1 - \sqrt[n]{1-CL} \qquad p \in [0, p_2]$$

Probabilità limite

Ho osservato x eventi:



- **limite** **inferiore** $p \in [p_1, 1]$**: se** $p < p_1$**, posso osservare** **almeno** x **eventi, ma in una frazione di esperimenti** $< 1 - CL$**.** **Se** $x = n$**,** $p_1^n = 1 - CL$**;**
- **limite** **superiore** $p \in [0, p_2]$**: se** $p > p_2$**, posso osservare** **fino a** x **eventi, ma in una frazione di esperimenti** $< 1 - CL$**.** **Se** $x = 0$**,** $(1 - p_2)^n = 1 - CL$**.**