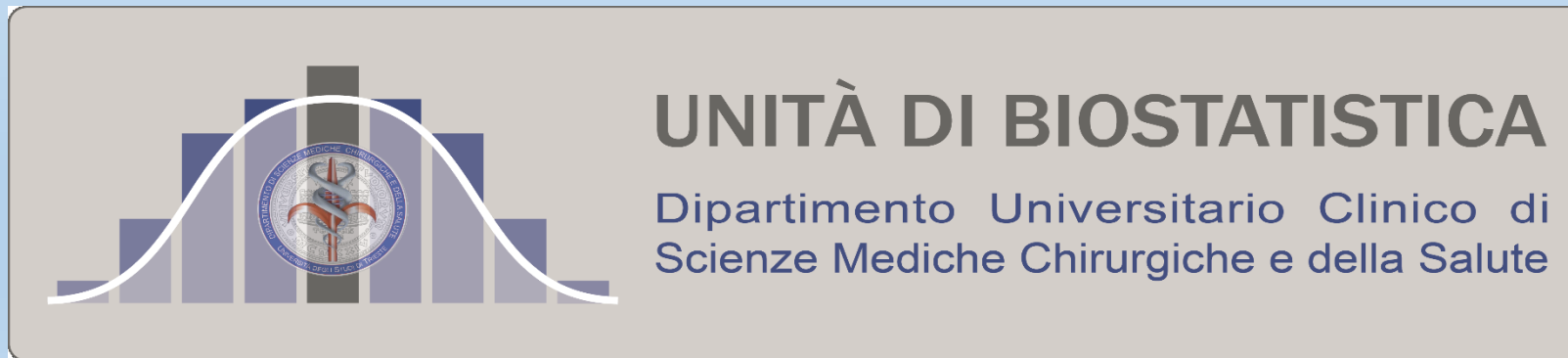


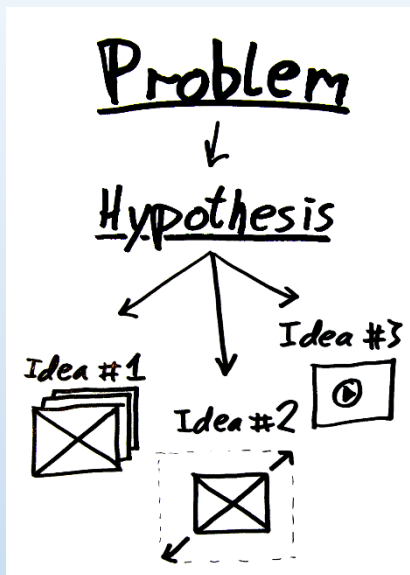
CDL in MEDICINA & CHIRURGIA

Statistica Medica

gbarbati@units.it

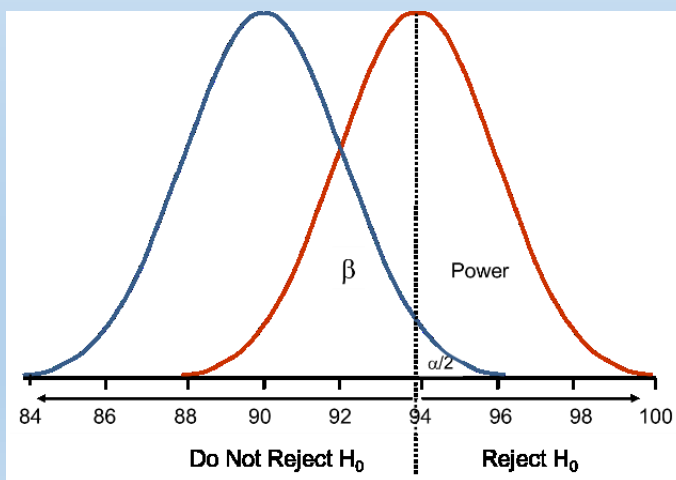
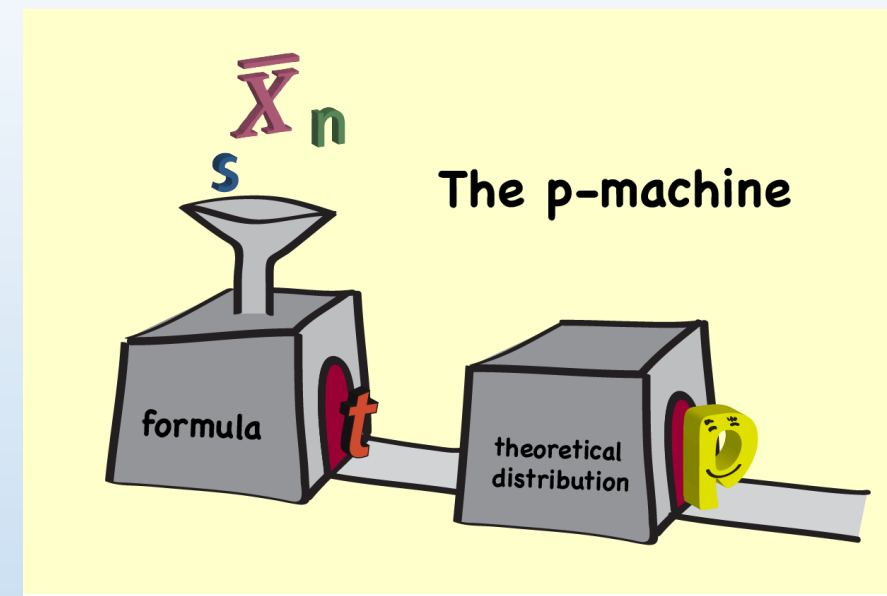
A.A. 2025-26



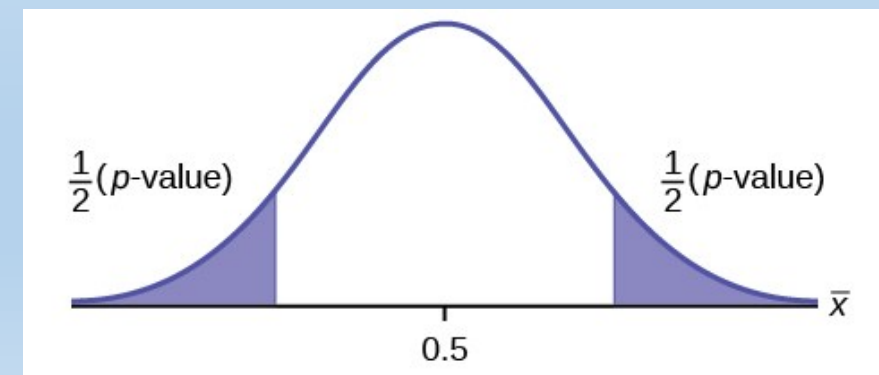


Sommario:

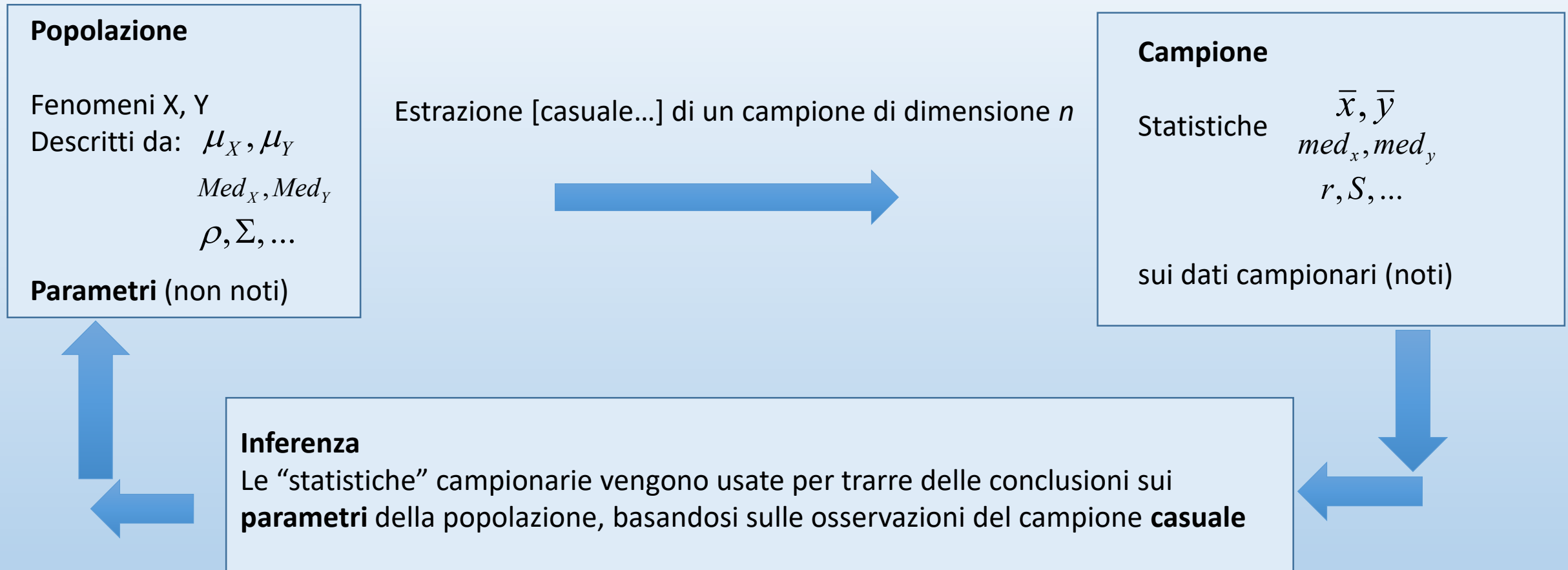
- Stima dei parametri
- Intervalli di confidenza
- Intervalli di credibilità



STATISTICS
MEAN NEVER
HAVING TO
SAY YOU'RE
CERTAIN



Dalla Statistica Descrittiva alla Statistica Inferenziale*



Introduciamo adesso il concetto di **stima di un parametro** tramite **campionamento**, utilizzando l'approccio **frequentista**.



R.A. Fisher 1890 – 1962

Fisher riteneva che la statistica dovesse essere **oggettiva**, basata esclusivamente sui dati osservati e non influenzata da opinioni o *credenze* soggettive.

Ripetibilità: Fisher sottolineava l'importanza della **ripetibilità** degli esperimenti.

Interpretazione: Fisher era molto critico nei confronti dell'interpretazione soggettiva delle probabilità. A suo avviso, la probabilità doveva essere intesa esclusivamente come **frequenza relativa** a lungo termine.

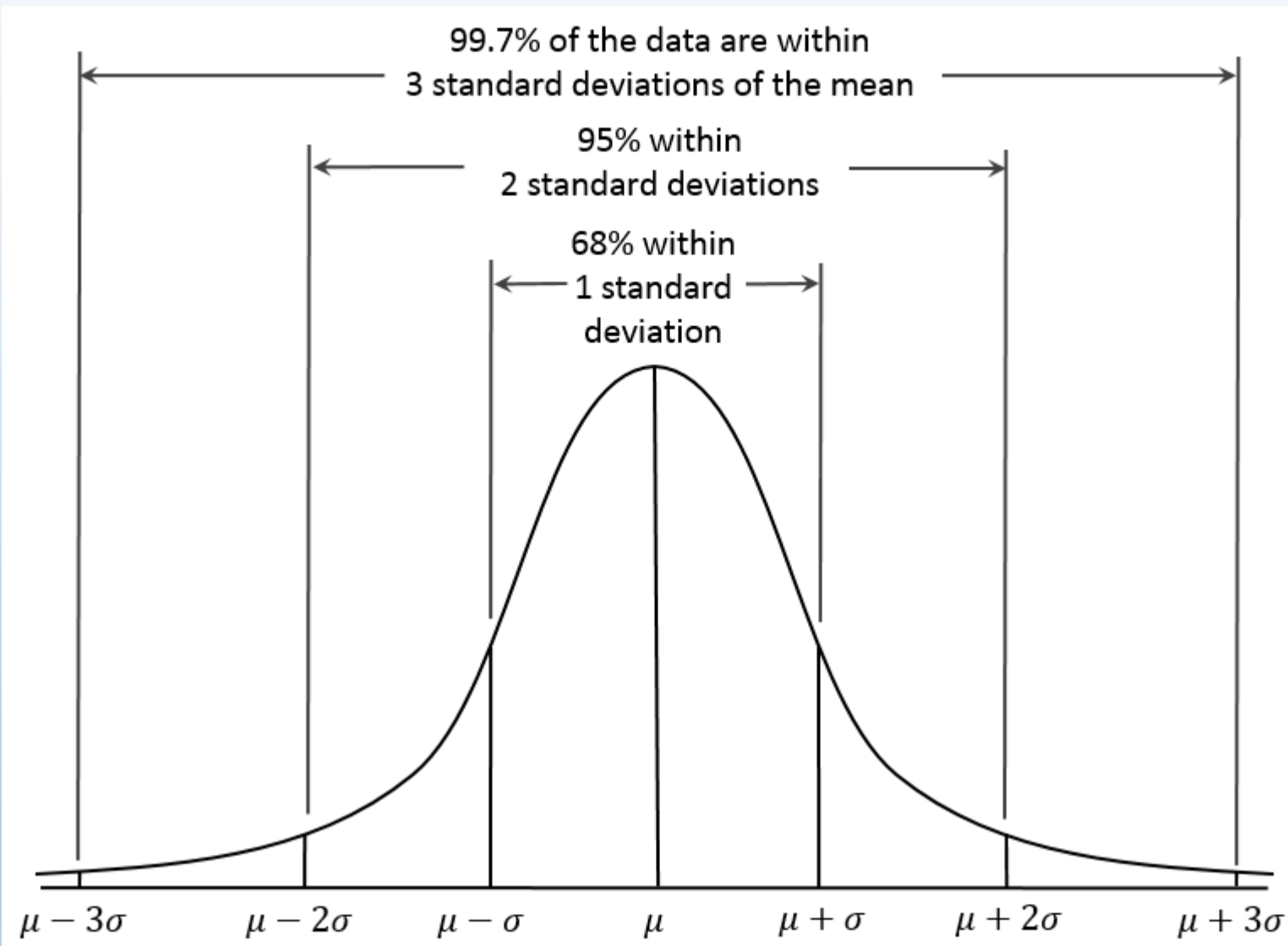
Parametri : sono costanti **fisse** ma sconosciute.

Le **probabilità** sono sempre interpretate come **frequenza relativa** a lungo termine.

Le procedure statistiche sono giudicate in base a come si comportano a lungo termine su un **numero infinito** di **ripetizioni ipotetiche** dell'esperimento (nelle stesse identiche condizioni)...

- **Intervalli di confidenza**
- **Test di ipotesi**





L'importanza della curva normale non risiede solo nella sua capacità di descrivere alcuni fenomeni su scala continua (*«forma» della distribuzione che andrebbe sempre verificata nei dati...*).

Essa occupa un posto di rilievo nella teoria dell'inferenza...

Gaussian distribution as **sum** of ...”others” random variables...

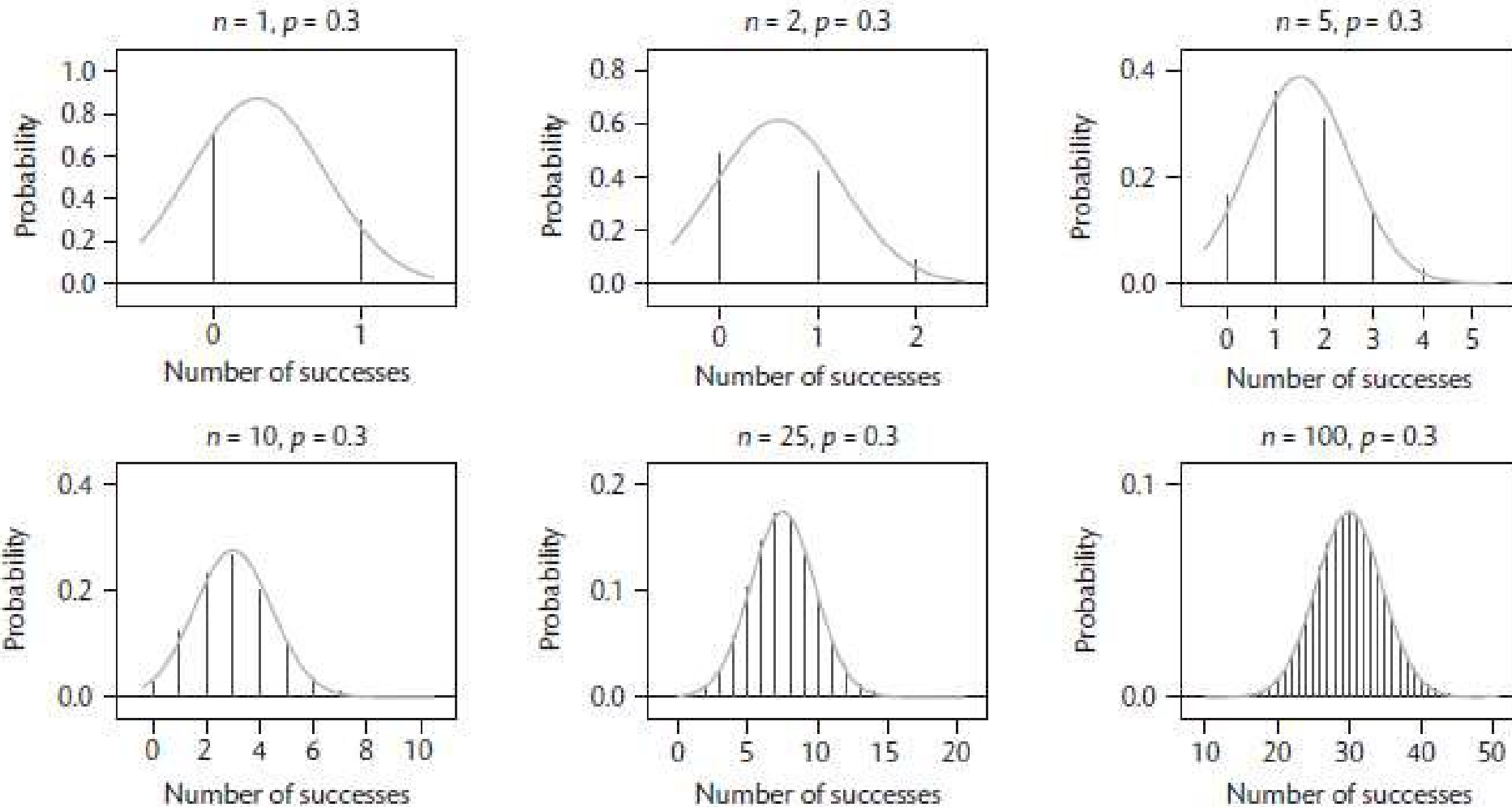
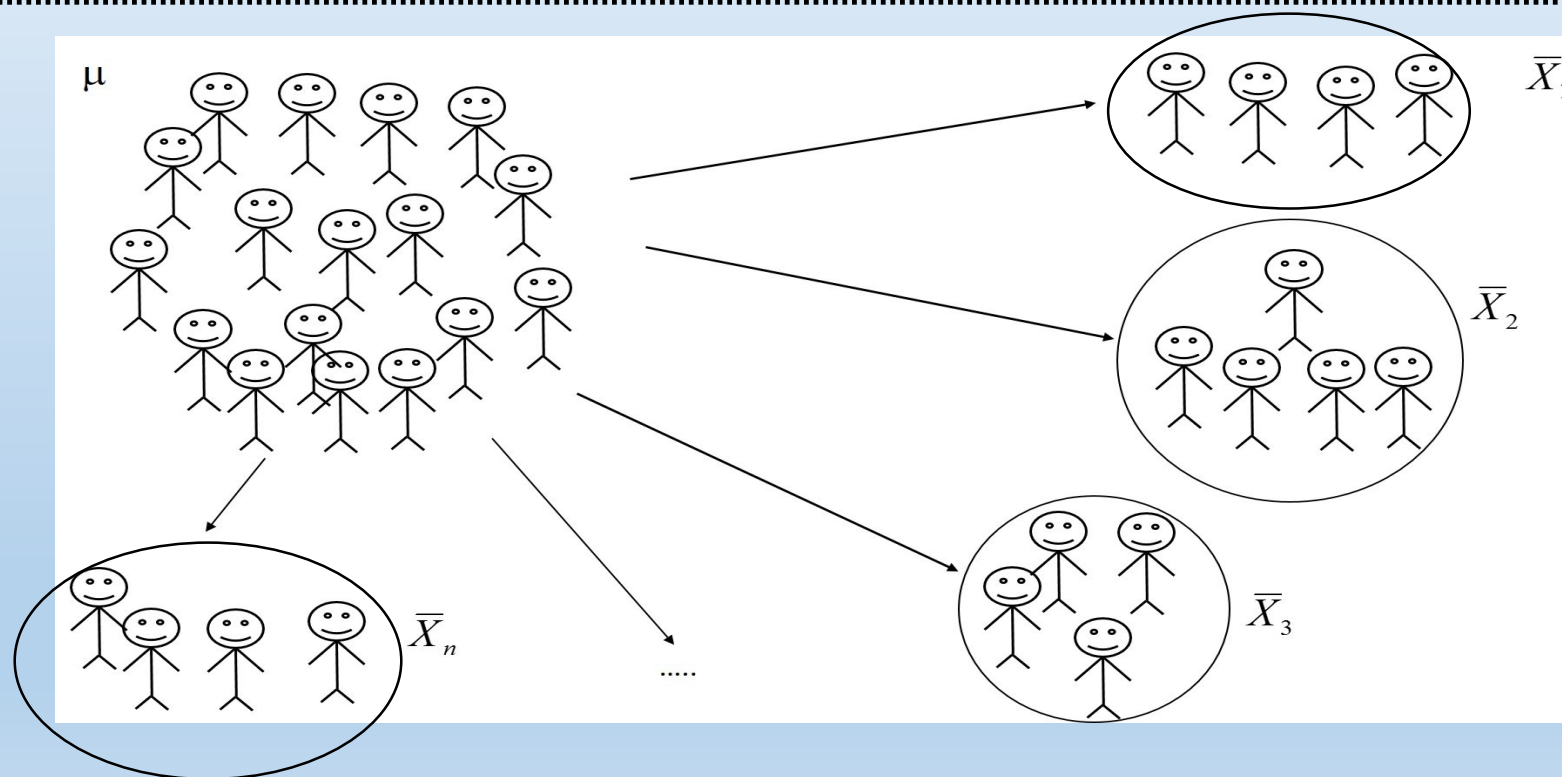
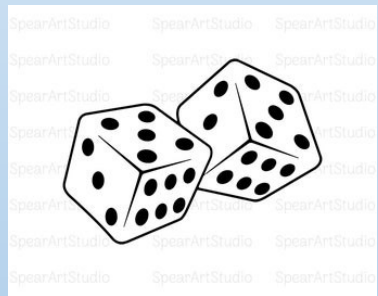


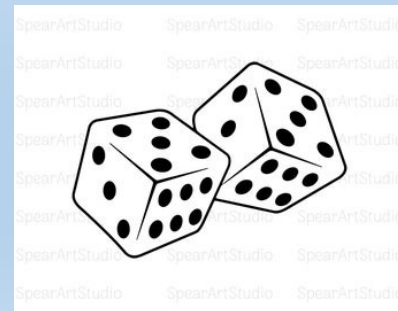
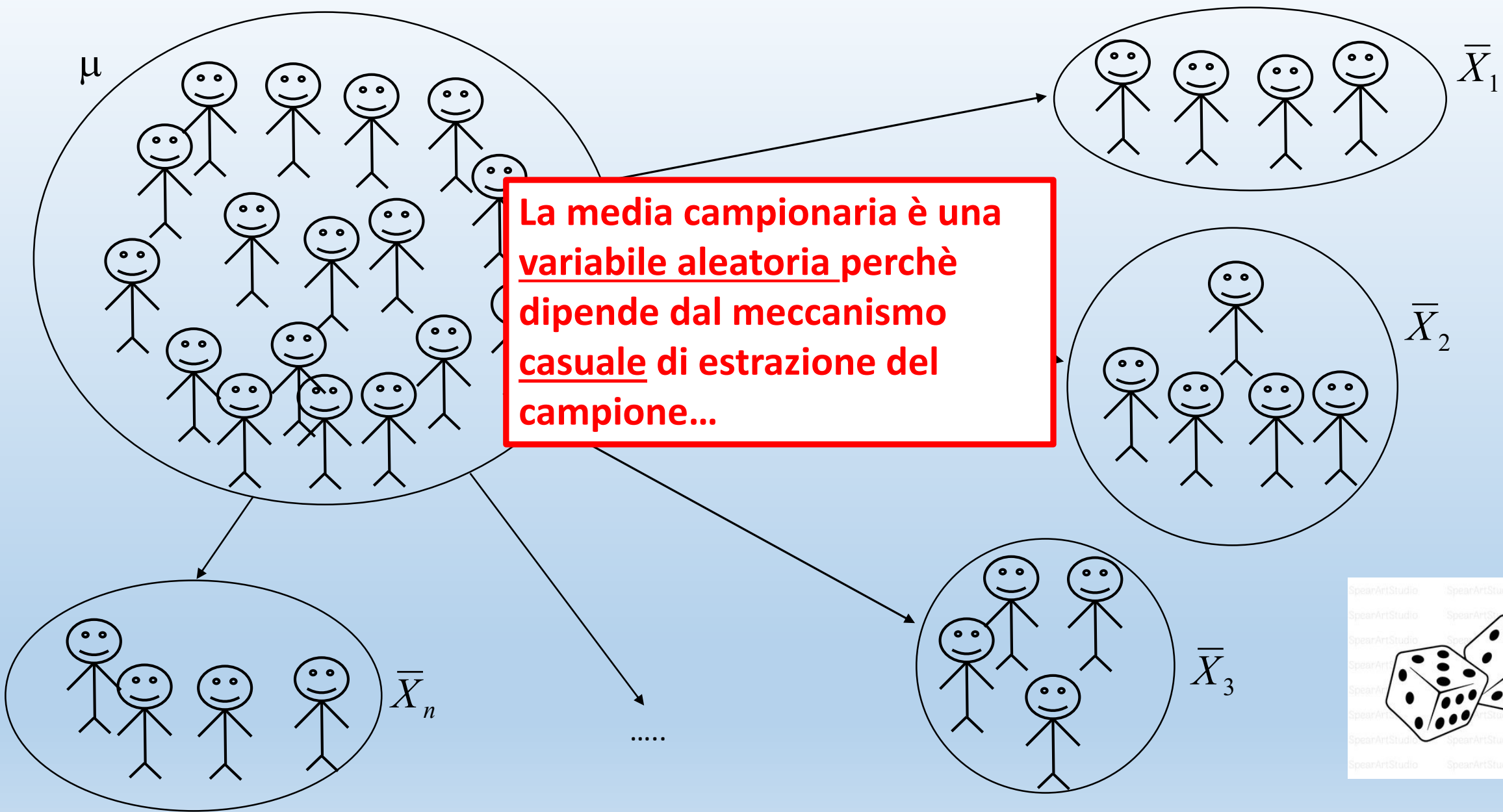
Figure 7.5 Binomial distributions for $p = 0.3$ and six different values of n , with corresponding Normal distribution curves.

Teorema del limite centrale

La **distribuzione di probabilità della somma (media)** di un numero *elevato** di variabili aleatorie indipendenti e *identicamente distribuite* tende distribuirsi **normalmente**, indipendentemente dalla distribuzione delle variabili originali.



* Un campione di dimensione $n > 30$ è considerato «sufficientemente grande»



Stima dei parametri

Obiettivo: Descrivere un fenomeno (su scala continua) oggetto del nostro studio.

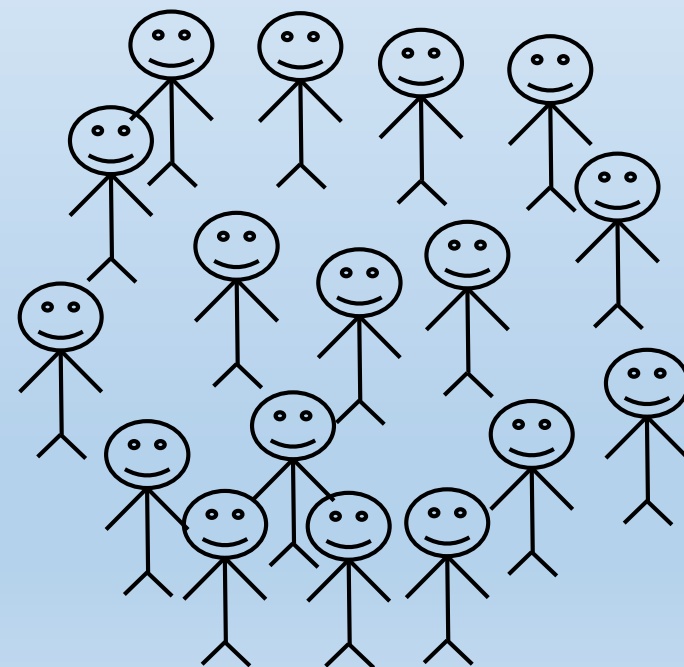
Es: Stimare il parametro μ del fenomeno nella popolazione (**ignoto**) tramite un campione di dimensione n

(anche la varianza del fenomeno σ^2 nella popolazione è **ignota**)

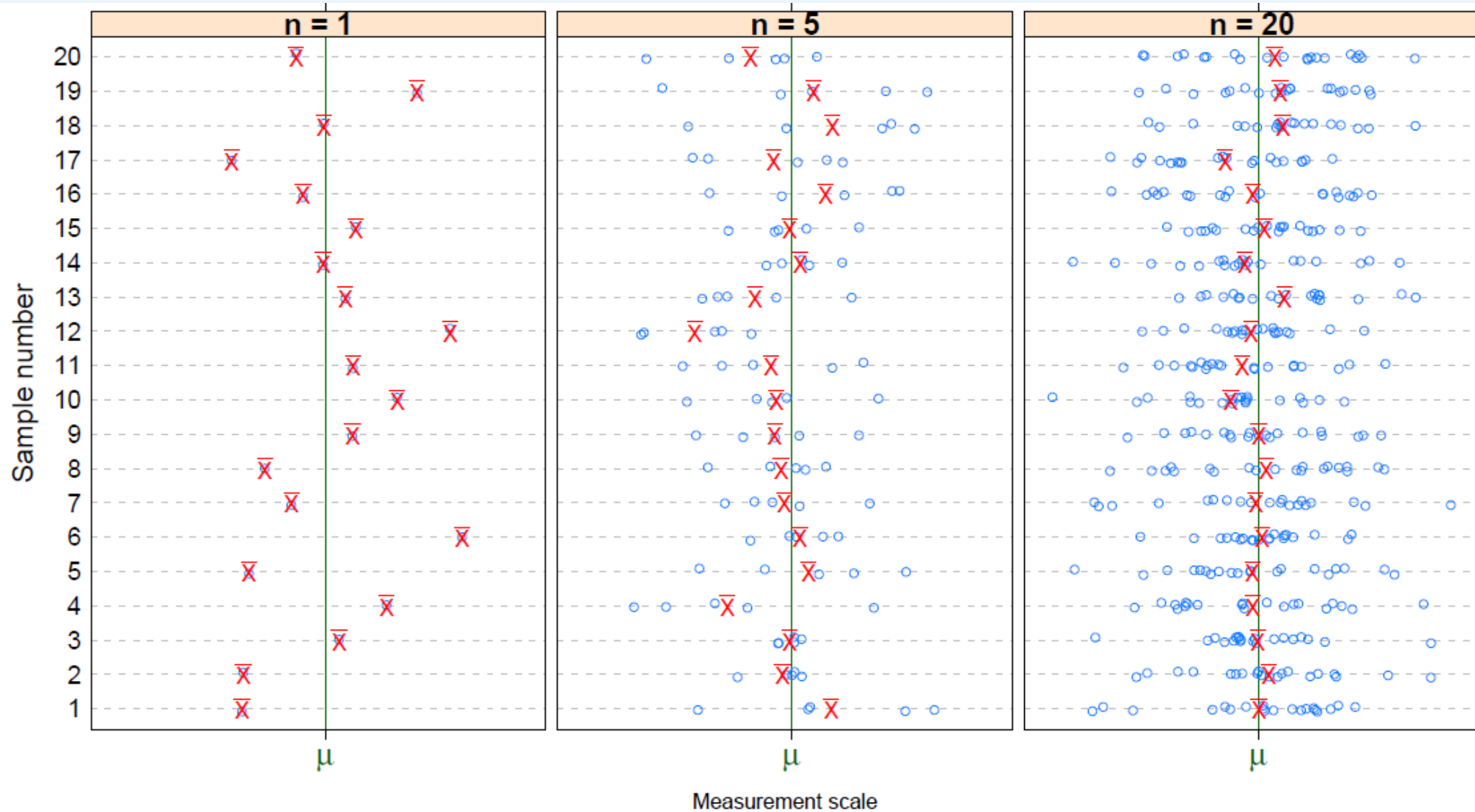
La (variabile aleatoria)
media campionaria $\bar{X}$ è un *buon* stimatore di μ

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

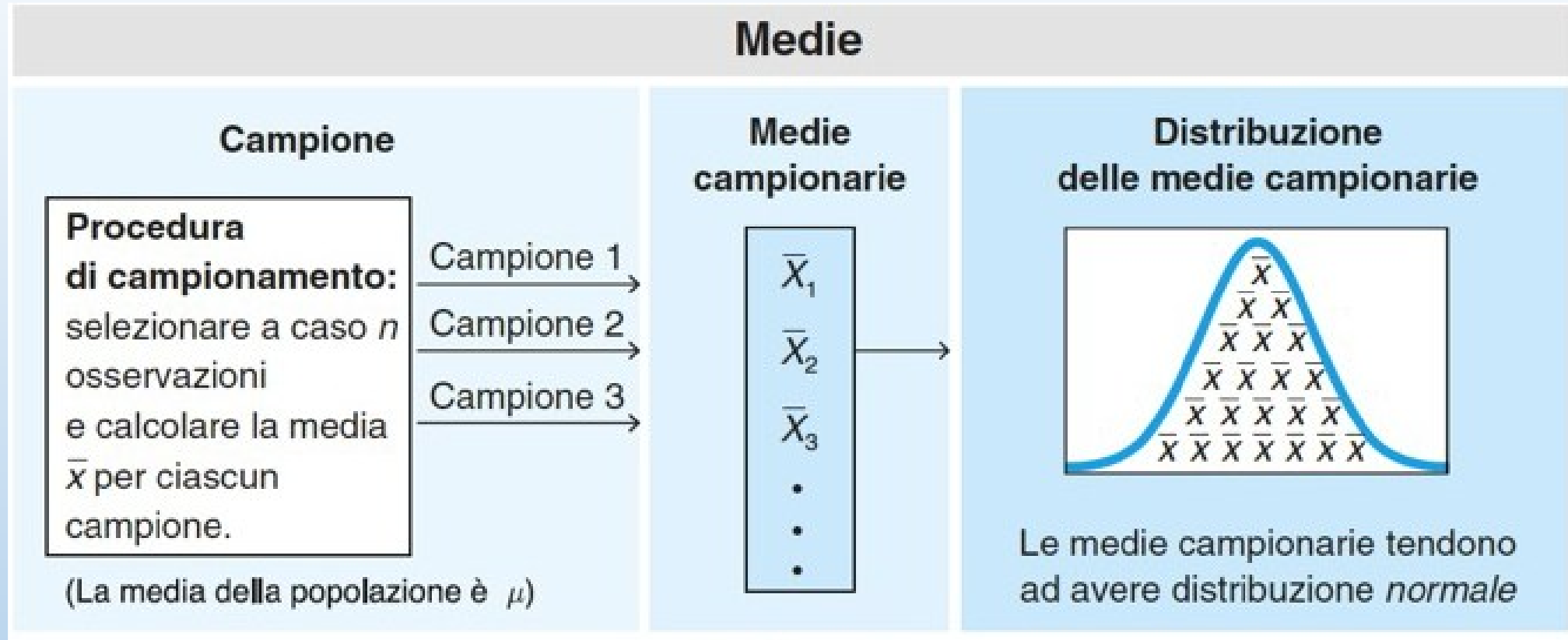
$$\bar{X} \rightarrow \mu \text{ per } n \rightarrow \infty$$



Simulazione della distribuzione della media campionaria



Distribuzione di **probabilità** della **media campionaria**



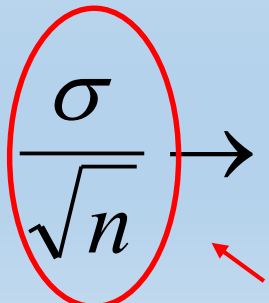
La **distribuzione campionaria** di una statistica è la distribuzione dei valori che la statistica può assumere quando consideriamo **tutti i possibili campioni di dimensione n** estratti dalla popolazione di partenza. La distribuzione campionaria di una statistica è rappresentata da una **distribuzione di probabilità**. Alcune distribuzioni campionarie sono ben note... ad esempio quella della media !

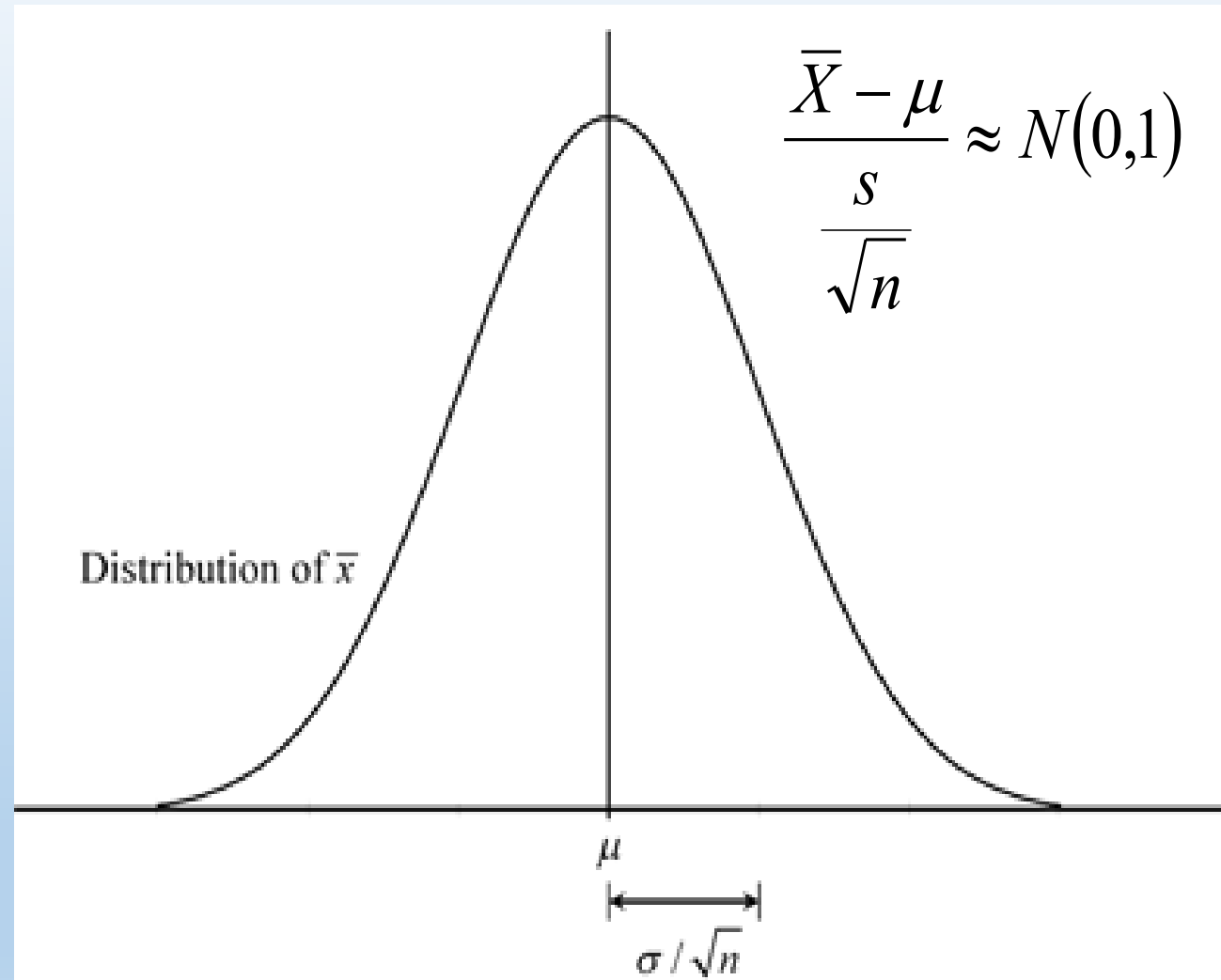
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{X} \rightarrow \mu \text{ per } n \rightarrow \infty$$

L' **errore** che commettiamo utilizzando la media campionaria come stima della media ignota è pari a:

$$devst(\bar{X}) = \frac{\sigma}{\sqrt{n}} \rightarrow 0 \text{ per } n \rightarrow \infty$$

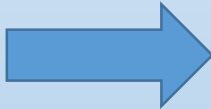

STANDARD ERROR



$\bar{X} \neq \mu$ praticamente sempre... anche se «*per caso*» $\bar{X} = \mu$...noi non lo sapremmo...

quanto è *sbagliata* $\bar{X}$ rispetto a μ ??

$$\text{devst}(\bar{X}) = \frac{\sigma}{\sqrt{n}} \rightarrow 0 \text{ per } n \rightarrow \infty$$

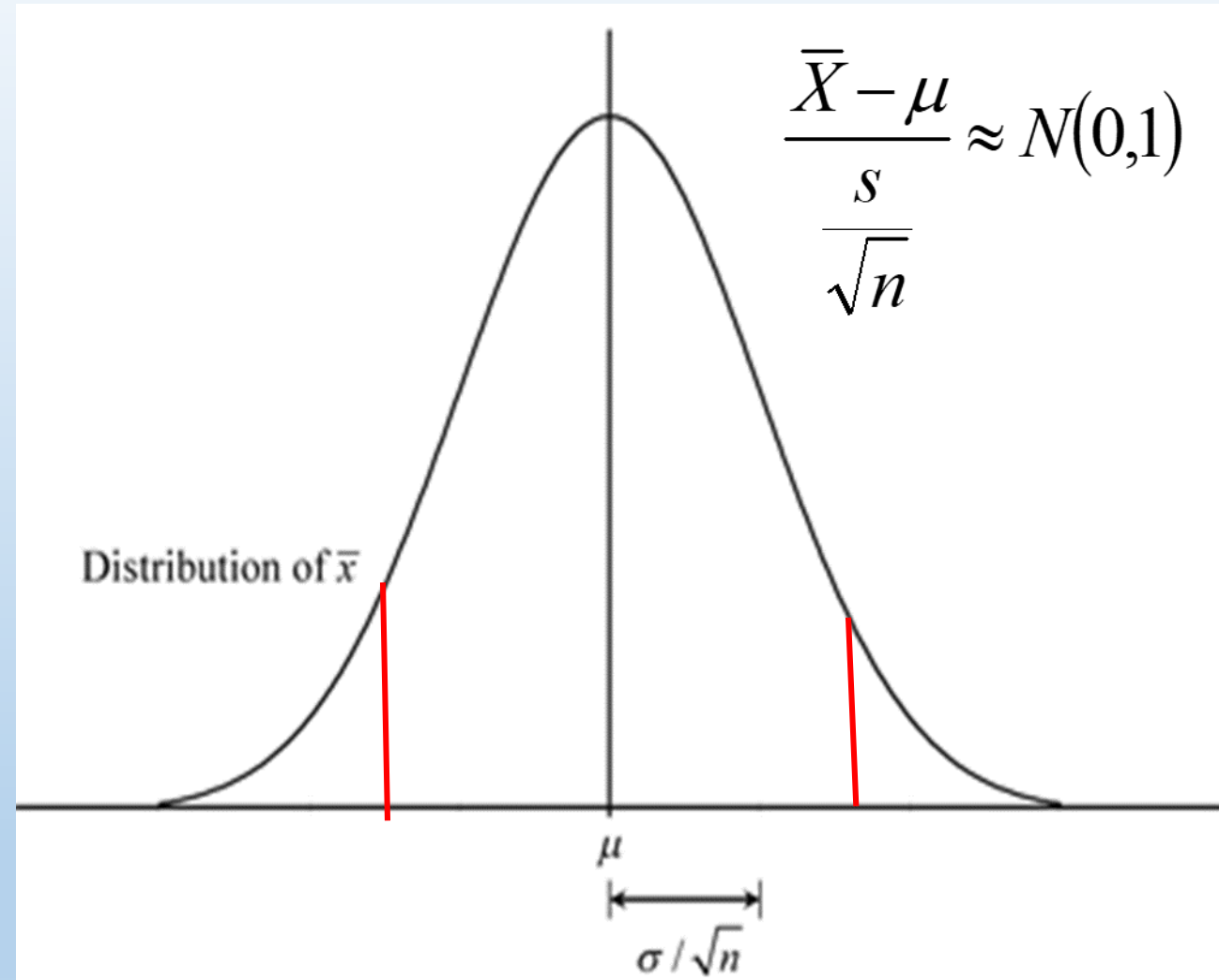
Ma la **varianza** della popolazione σ^2 non è nota  come stimare l'errore su $\bar{X}$?

s^2 (**deviazione standard campionaria**) è una **buona** stima di σ^2 :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Ricapitolando:

- la **distribuzione di probabilità** della media campionaria è approssimata da una gaussiana *centrata* sul parametro ignoto μ
- usando le proprietà della gaussiana, possiamo calcolare degli **intervalli** che contengono una certa % dei dati della distribuzione (ad esempio il **95%** dei valori intorno alla media)...



$$P(\text{LOWER LIMIT} < \mu < \text{UPPER LIMIT}) = 0.95 = (1-\alpha)$$

$$P(\text{LOWER LIMIT} < \mu < \text{UPPER LIMIT}) = 0.95 = (1-\alpha)$$

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

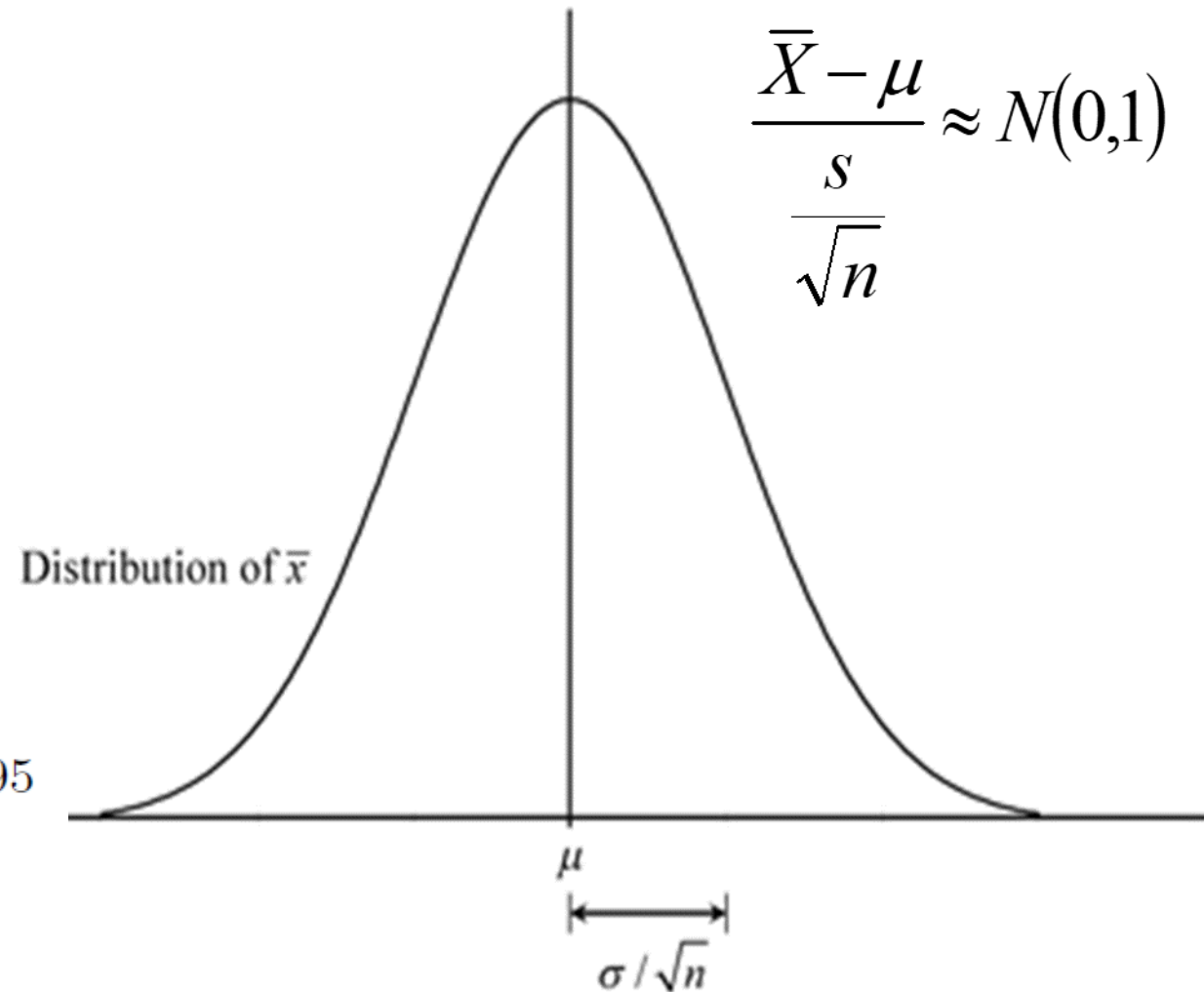
$$P\left(-1.96 \frac{\sigma}{\sqrt{n}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \frac{\sigma}{\sqrt{n}} \leq 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-1.96 \frac{\sigma}{\sqrt{n}} - \bar{X} \leq \bar{X} - \mu - \bar{X} \leq 1.96 \frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 0.95$$

$$P\left(1.96 \frac{\sigma}{\sqrt{n}} + \bar{X} \geq \mu \geq -1.96 \frac{\sigma}{\sqrt{n}} + \bar{X}\right) = 0.95$$

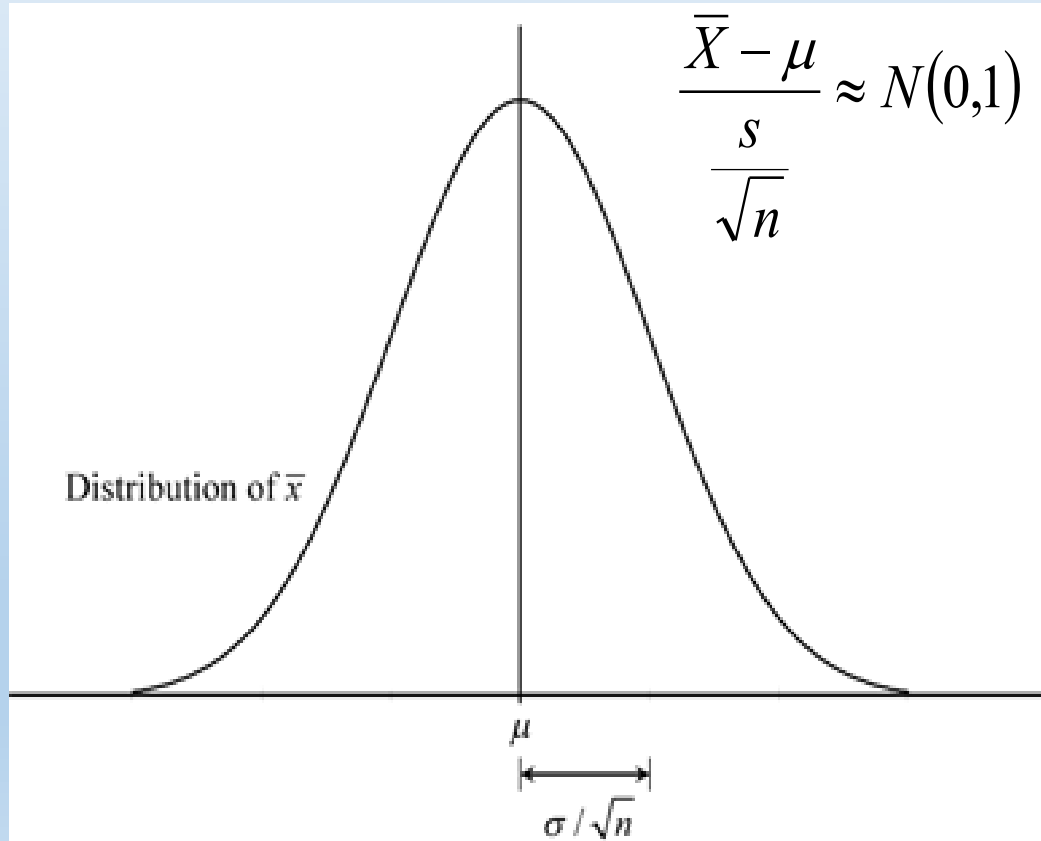
$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$



INTERVALLO DI CONFIDENZA (IC)

$$\left[\bar{X} - \text{cost} * \frac{s}{\sqrt{n}}; \bar{X} + \text{cost} * \frac{s}{\sqrt{n}} \right]$$

Questo **intervallo** *contiene* il parametro μ con una probabilità determinata dalla costante utilizzata (**nel lungo termine**)...



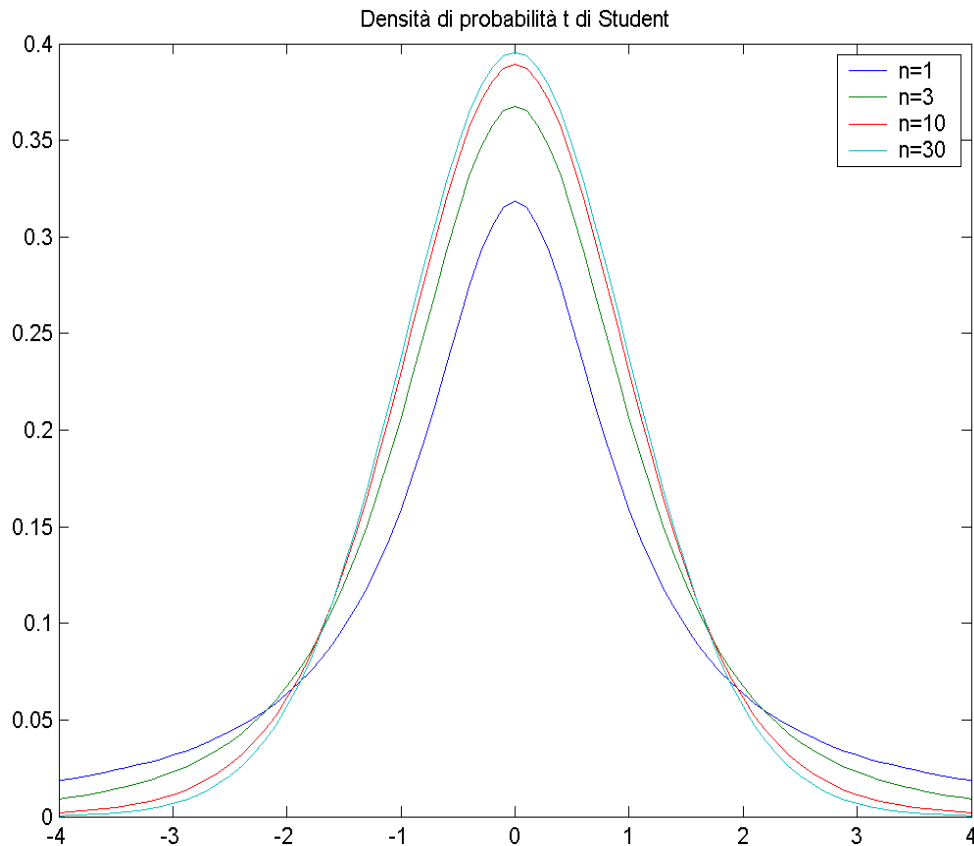
Se vogliamo un IC al 95%:

$$\left[\bar{X} - 1.96 * \frac{s}{\sqrt{n}}; \bar{X} + 1.96 * \frac{s}{\sqrt{n}} \right]$$

Standard Error : SE

Correzione per *n* «piccolo»: distribuzione t di Student

Per una migliore stima dell'IC in piccoli campioni si definisce la VA di Student (t di Student):
$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$



La distribuzione **t di Student** è molto simile alla gaussiana. Ma cambia di *forma* in relazione alla numerosità *n* del campione: tende ad avvicinarsi alla **distribuzione normale standard** $N(0,1)$, al crescere di *n*.

Per $n > 30$ le due distribuzioni sono indistinguibili.

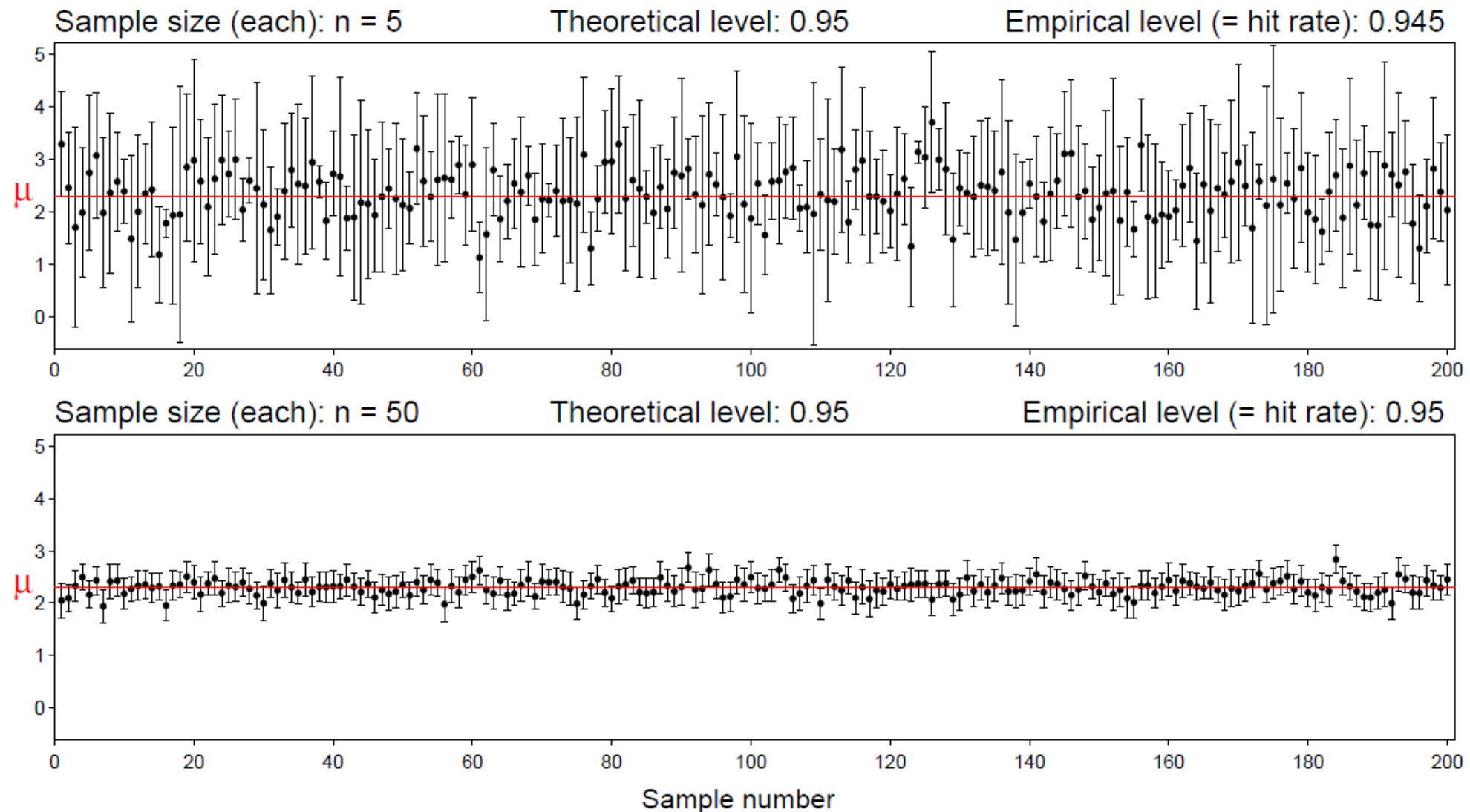
IC al 95%:

$$\left[\bar{X} - t_n * \frac{s}{\sqrt{n}} ; \bar{X} + t_n * \frac{s}{\sqrt{n}} \right]$$



William Gosset, detto "Student" (1876-1937)

«Hit rate» dell'IC di Student per 2 x 200 campioni simulati



Confrontiamo queste due espressioni:

1. L'intervallo di confidenza **contiene** il parametro ignoto μ **con una probabilità del X%...**
2. **Il parametro ignoto μ ha una probabilità del X% di cadere in quell'intervallo...**

Sembrano simili, ma nell'impostazione frequentista la seconda è sbagliata.

Il parametro μ **NON E'** una variabile aleatoria; è un valore unico e non noto.

Per ogni specifico campione estratto abbiamo solo 2 possibilità: o μ è dentro l'intervallo o è fuori dall'intervallo.

Stiamo facendo riferimento ***ad una serie ipotetica di campionamenti ripetuti*** e stiamo dicendo che nel X% dei casi gli intervalli di confidenza calcolati ***conterranno*** μ . Nel restante 5% non lo conterranno.

l'impostazione **bayesiana** che deriva dalla definizione soggettiva di probabilità cambia le carte in tavola...

Parametri : sono costanti **fisse** ma sconosciute.

Le **probabilità** sono sempre interpretate come **frequenza relativa** a lungo termine.

Le procedure statistiche sono giudicate in base a come si comportano a lungo termine su un *numero infinito* di *ripetizioni ipotetiche* dell'esperimento.



- Intervalli di confidenza
- Test di ipotesi

IMPOSTAZIONE FREQUENTISTA

Poiché siamo incerti sul valore vero dei **parametri**, li considereremo come **variabili casuali**.

Le regole della probabilità vengono utilizzate *direttamente* per fare inferenze sui parametri.

Probabilità a priori + dati => Probabilità a posteriori



- Intervalli di credibilità
- Fattore di Bayes

IMPOSTAZIONE BAYESIANA

INTERVALLO DI CREDIBILITA' [CENNI]

Il parametro *ignoto* ha una probabilità del X% di cadere in quell'intervallo...

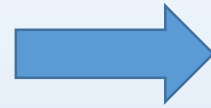
Intervallo di credibilità: *"C'è il 95% di probabilità che il valore sia in questo intervallo"*.

Intervallo di confidenza: *"Se ripetessimo l'esperimento molte volte, il 95% degli intervalli calcolati conterrebbe il vero valore"*.

COME SI COSTRUISCE UN INTERVALLO DI CREDIBILITA' ?

1. Si definisce la **distribuzione a priori**: si sceglie una distribuzione che rifletta la **conoscenza iniziale** sul parametro.
2. Si calcola la **verosimiglianza**: la probabilità di *osservare* i dati condizionata ai possibili valori del parametro.
3. Si calcola la **distribuzione a posteriori**: utilizzando il teorema di Bayes si combina la distribuzione a priori e la verosimiglianza per ottenere la distribuzione a posteriori.
4. Si determina la **stima puntuale** e i **quantili**: indicatori di posizione / quantili della **distribuzione a posteriori** che delimitano l'intervallo di credibilità desiderato.

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)}$$



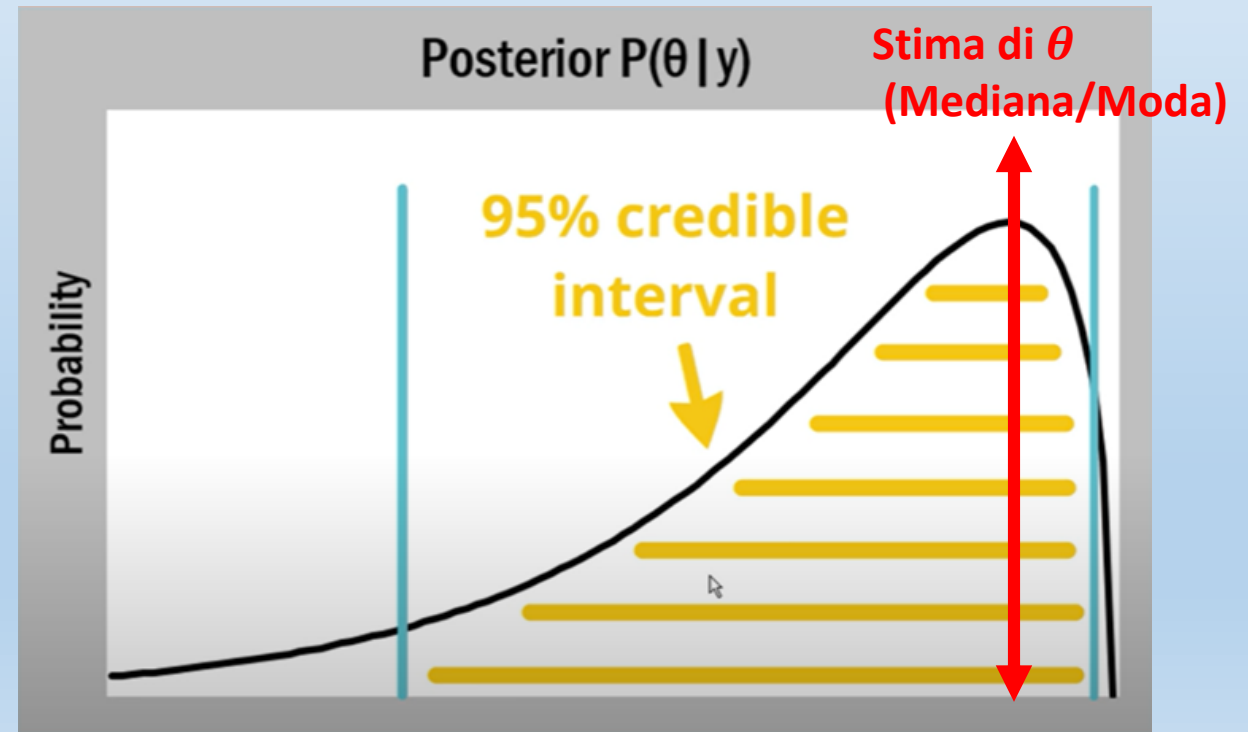
$$P(\theta|data) \cong P(data|\theta)P(\theta)$$

$$Posterior \cong Likelihood * Prior$$

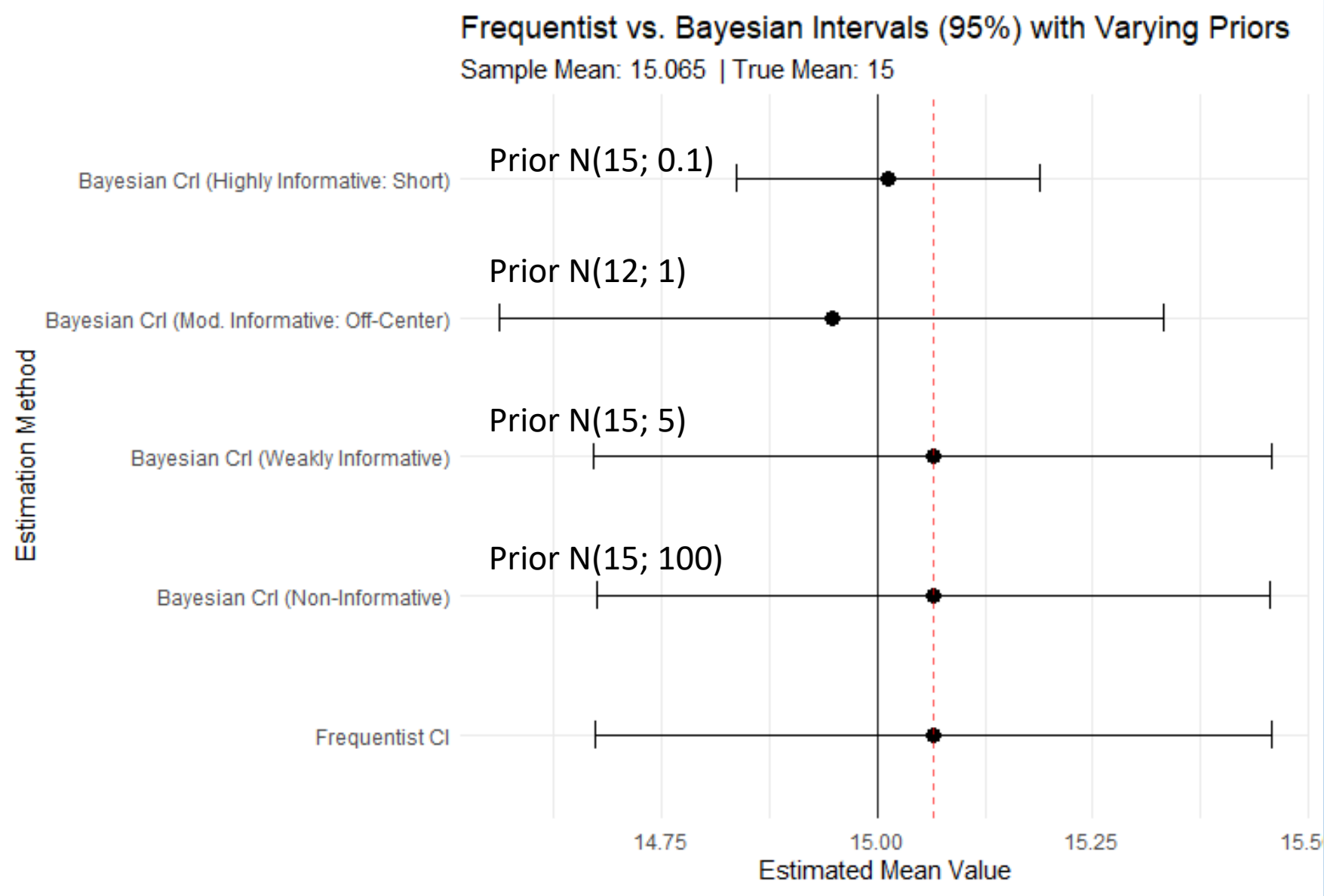
bilanciamento tra i dati e il prior: al crescere della dimensione campionaria «cresce» il peso dei dati

In base alla distribuzione a priori che scegliamo e alla probabilità di osservare i dati in funzione dei possibili valori di θ deriviamo la distribuzione a posteriori. Da questa, deriviamo la stima puntuale di θ e l'intervallo di credibilità.

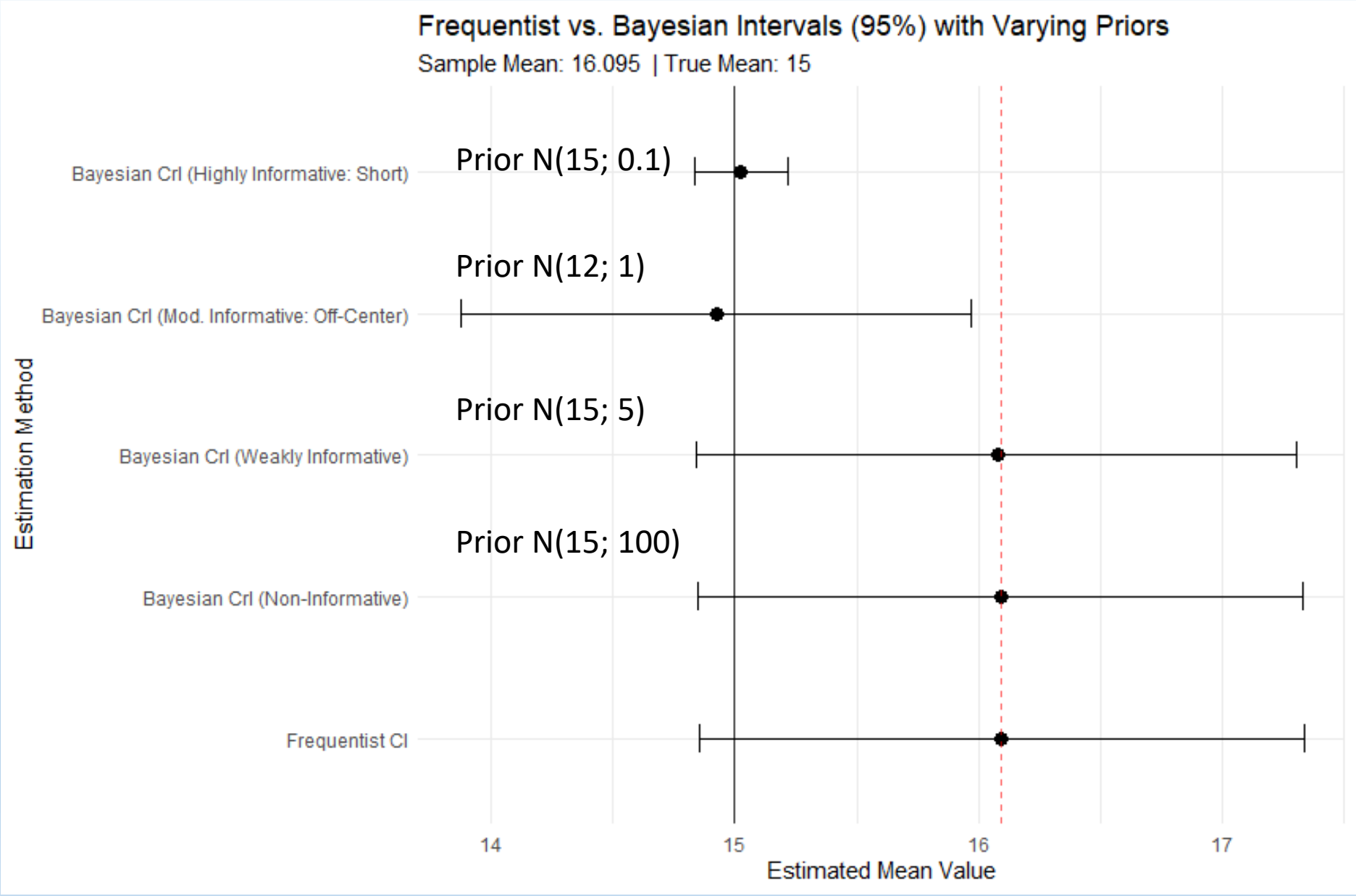
N.B: la distribuzione del *Prior* viene ricavata da studi precedenti/ipotesi...



*Dati generati da una $N(15; 2)$, campione di numerosità 100



*Dati generati da una $N(15; 2)$, campione di numerosità 10



VA Binomiale (RECAP)

Una VA X Binomiale «conta» il numero k dei successi di n VA di Bernoulli.

$$Y = X_1 + X_2 + \dots + X_n \qquad Y \approx \text{Bin}(n, p)$$

Es: Quale è la probabilità che in n lanci di una moneta esca k volte testa?

Due parametri descrivono la distribuzione di X :

- il numero delle prove n
- la probabilità di successo ad ogni prova p



$$P(Y = k) = \frac{n!}{(n-k)!k!} p^k q^{n-k}$$

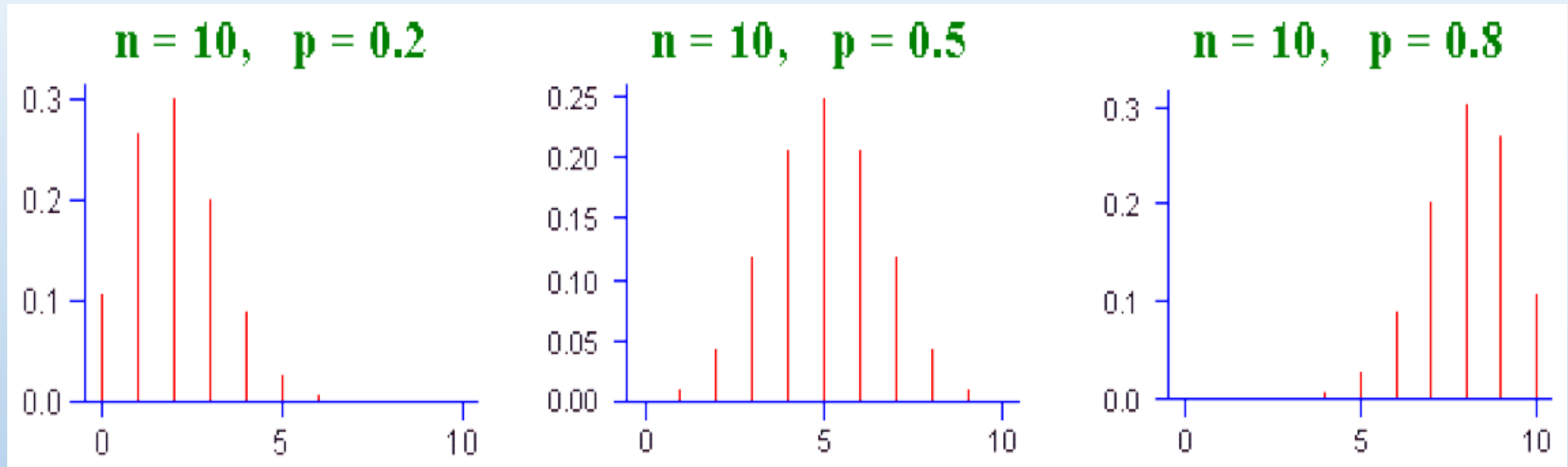


Coefficiente binomiale:

Conta tutti i possibili modi in cui possiamo estrarre k elementi da n

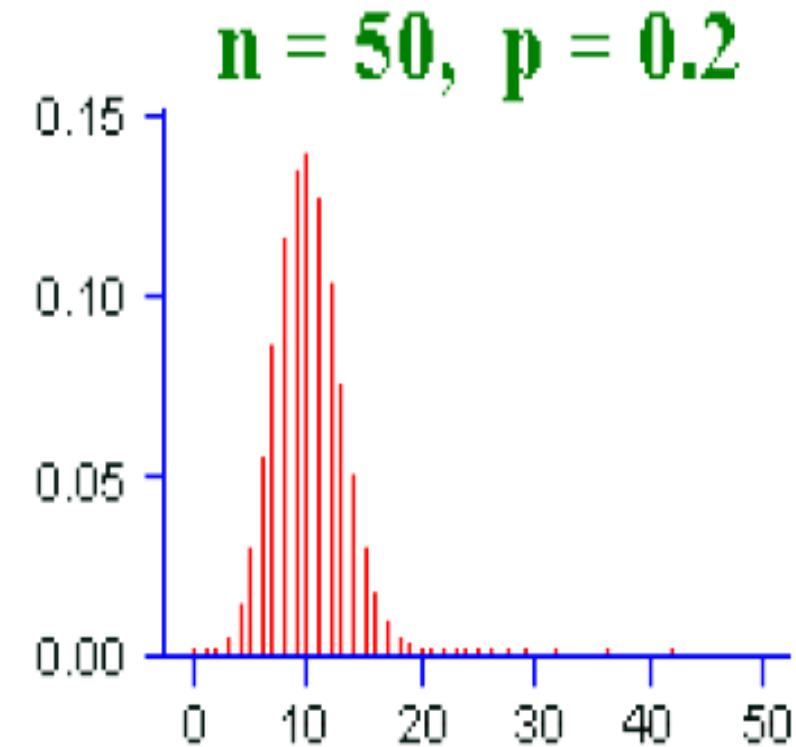
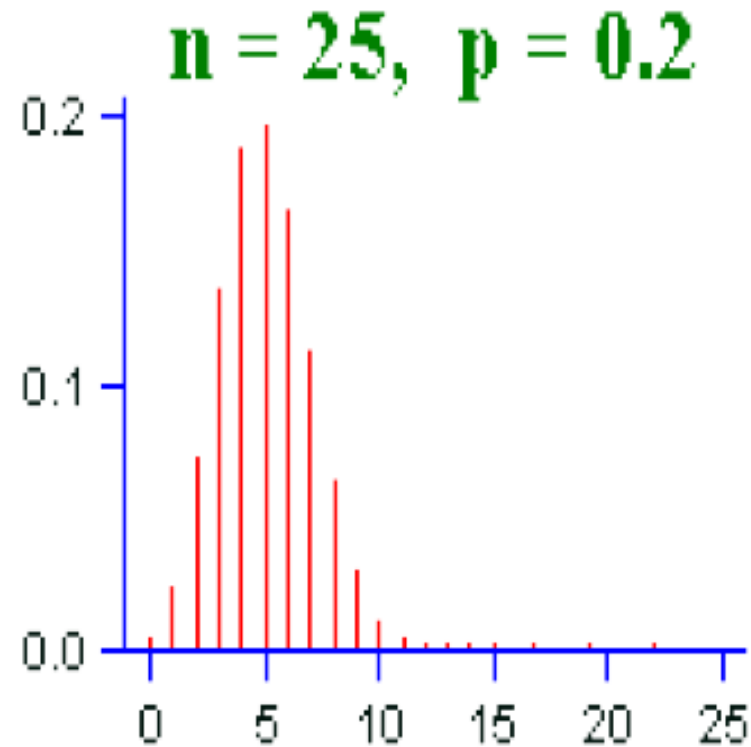
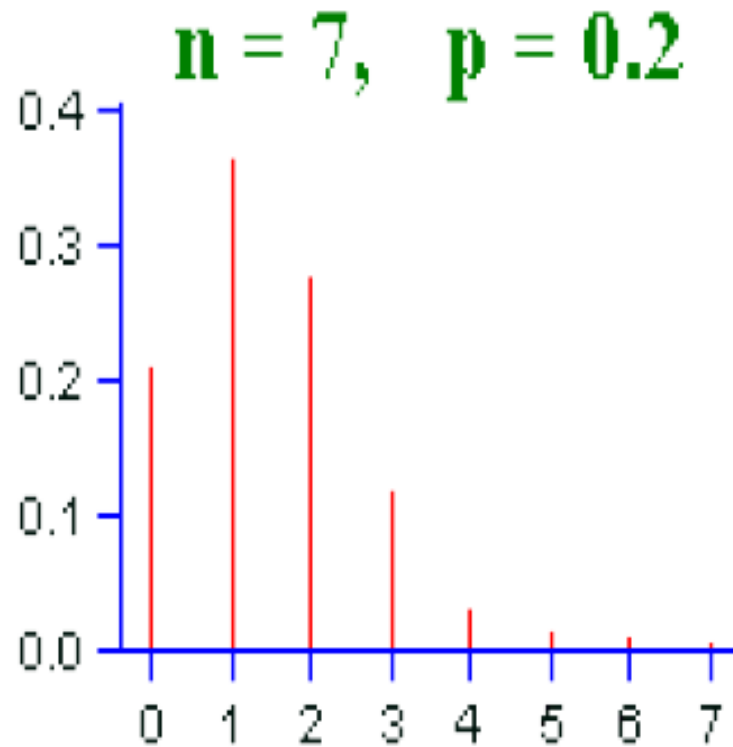
$$\longrightarrow \binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{1*2*3\dots*k} = \frac{n!}{(n-k)!k!}$$

La “**forma**” della distribuzione di probabilità una VA binomiale dipende dai valori di n e p :



- Quando p è piccolo (0.2), la distribuzione binomiale è asimmetrica a destra
- Quando $p = 0.5$, la distribuzione binomiale è simmetrica
- Quando $p > 0.5$, la distribuzione binomiale è asimmetrica a sinistra...

Distribuzioni binomiali per diversi valori di n , dato $p=0.2$:



Al crescere di n , la distribuzione binomiale diventa sempre più simmetrica..

La distribuzione binomiale **può essere approssimata da una distribuzione normale** (per qualsiasi valore di p) al crescere di n^* .

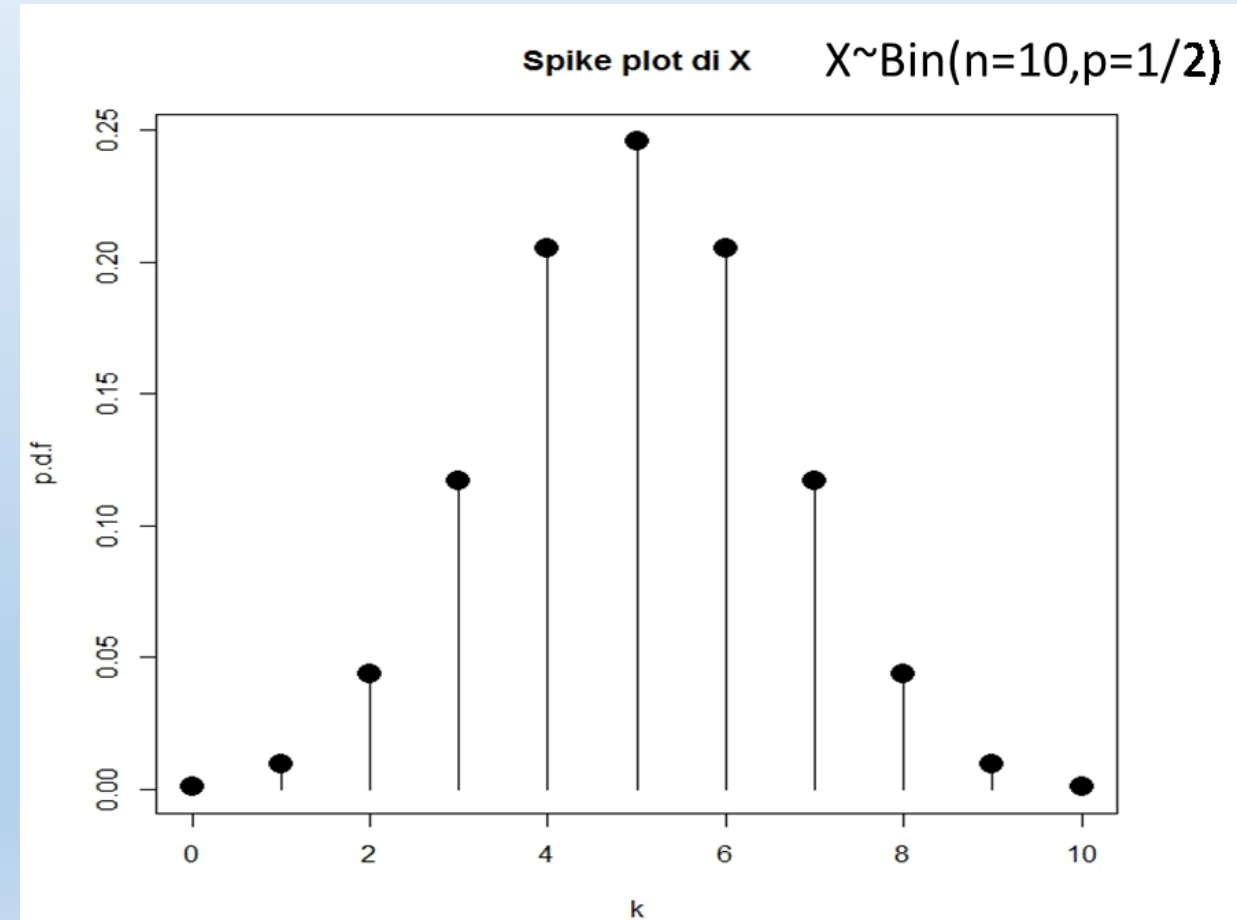
Questa approssimazione è accettabile data la condizione: $np \geq 5$ e $n(1 - p) \geq 5$

$$Y \approx \text{Bin}(n, p)$$

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} \rightarrow N(0,1)$$

La media di una VA Binomiale è: $\mu = np$

La deviazione standard è: $\sigma = \sqrt{np(1-p)}$



Distribuzione di probabilità della proporzione campionaria

Proporzioni

Campione

Procedura di campionamento: selezionare a caso n osservazioni e calcolare la proporzione $\hat{p}$ per ciascun campione.

(La proporzione della popolazione è p)

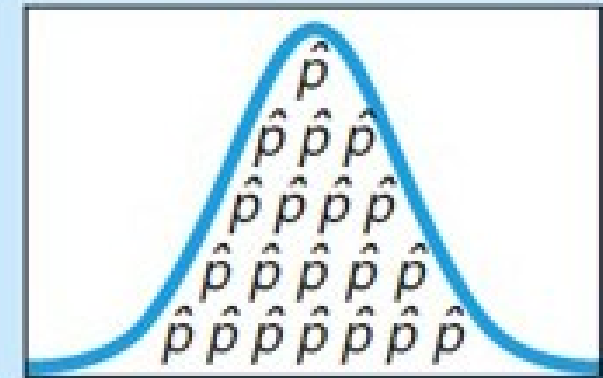
Campione 1 →
Campione 2 →
Campione 3 →

Proporzioni campionarie

$\hat{p}_1$
 $\hat{p}_2$
 $\hat{p}_3$
•
•
•



Distribuzione delle proporzioni campionarie



Le proporzioni campionarie tendono ad avere distribuzione *normale*

Intervallo di **confidenza** per una proporzione

Stima di una proporzione nella popolazione: da un campione casuale di **VA** $X_1, X_2, \dots, X_n$ che possono assumere solo due possibili valori 0 o 1 (sano o malato...) siamo interessati a stimare la proporzione di eventi (=malati) nella popolazione: $p = P(X_i=1)$.

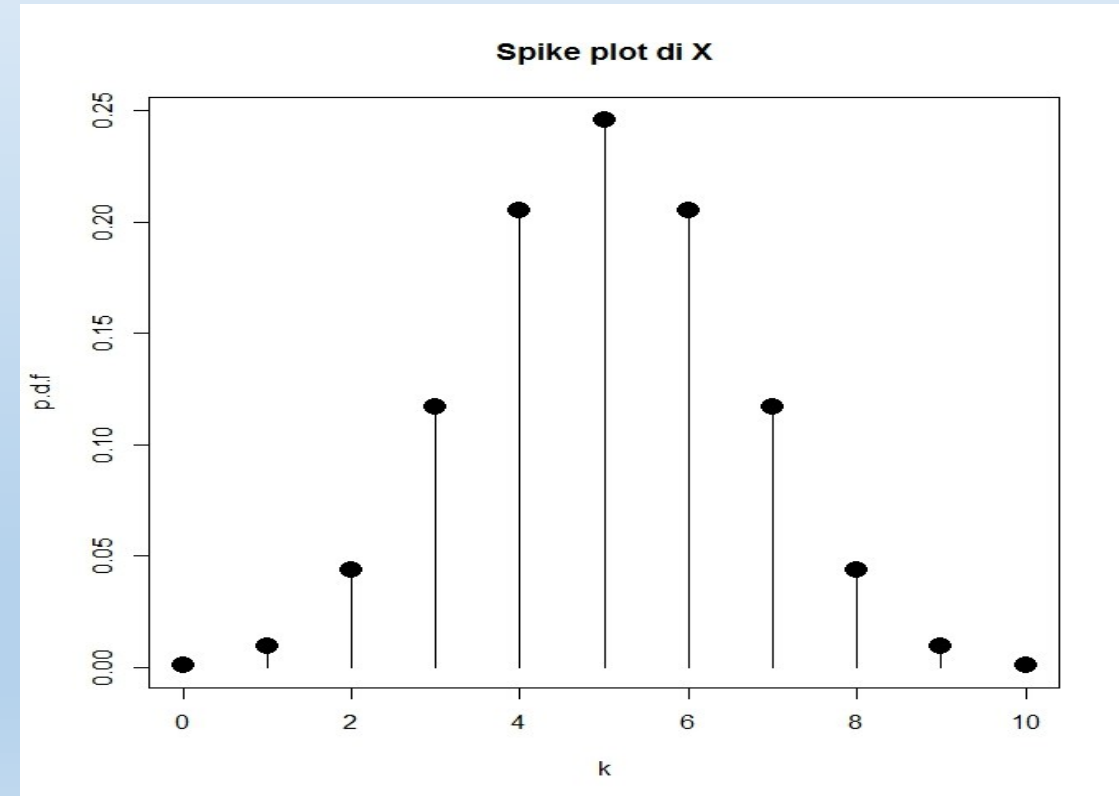
(ben approssimata dalla gaussiana per $np \geq 5$ e $n(1-p) \geq 5$)

$$\bar{p} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{stimatore di } p$$

$$\frac{\bar{p} - p}{SE(\bar{p})} \approx N(0,1)$$

$$\bar{p} \pm 1.96 * SE(\bar{p})$$

$$SE(\bar{p}) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad \text{errore standard di } \bar{p}$$



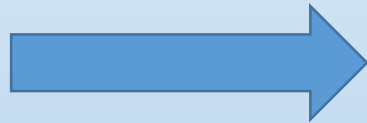
Intervallo di confidenza per una proporzione: esempio

Siamo interessati a stimare la proporzione di studenti universitari che dormono almeno 8 ore a notte.

Analizziamo un campione di 27 studenti, dei quali 11 ci dicono di dormire 8 ore o più x notte (41%).

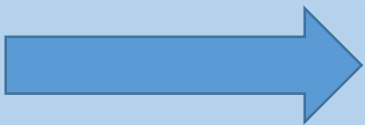
$$X \sim \text{Bin}(27, p)$$

$$\bar{p} \pm 1.96 * SE(\bar{p})$$



N = 27 Sample size
x = 11 Number in the sample with the result or finding in question
CL = 95 % Confidence level

Calculate



Proportion of positive results = $P = x/N = 0.4074$
Lower bound = $P - (Z_{\alpha} * SEM) = 0.2221$
Upper bound = $P + (Z_{\alpha} * SEM) = 0.5927$

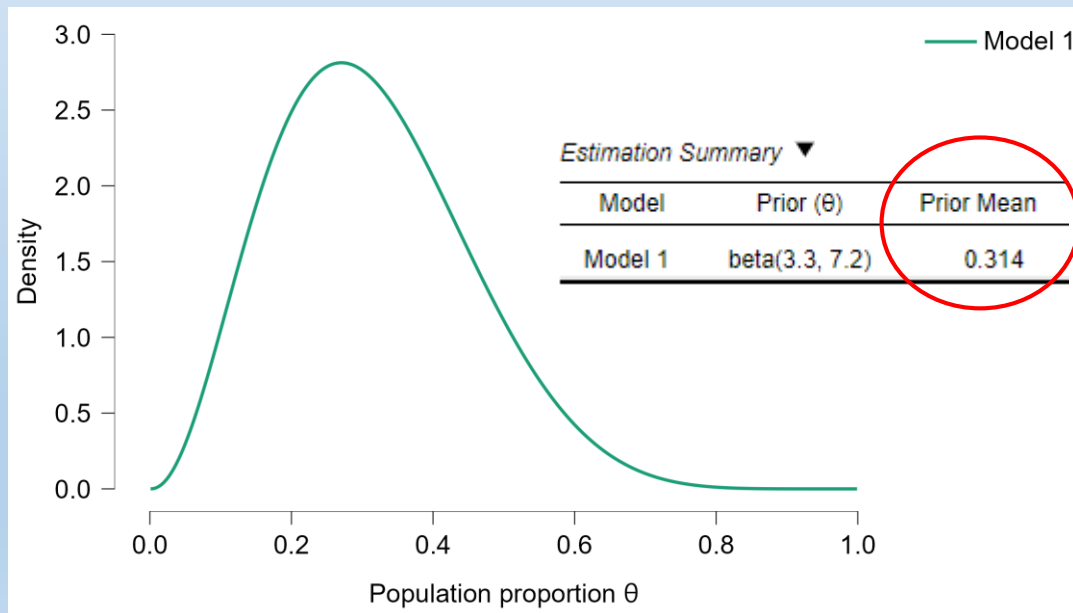
Intervallo di credibilità per una proporzione: esempio

Siamo interessati a stimare la proporzione di studenti universitari che dormono almeno 8 ore a notte.

Analizziamo un campione di 27 studenti, dei quali 11 ci dicono di dormire 8 ore o più x notte (41%).

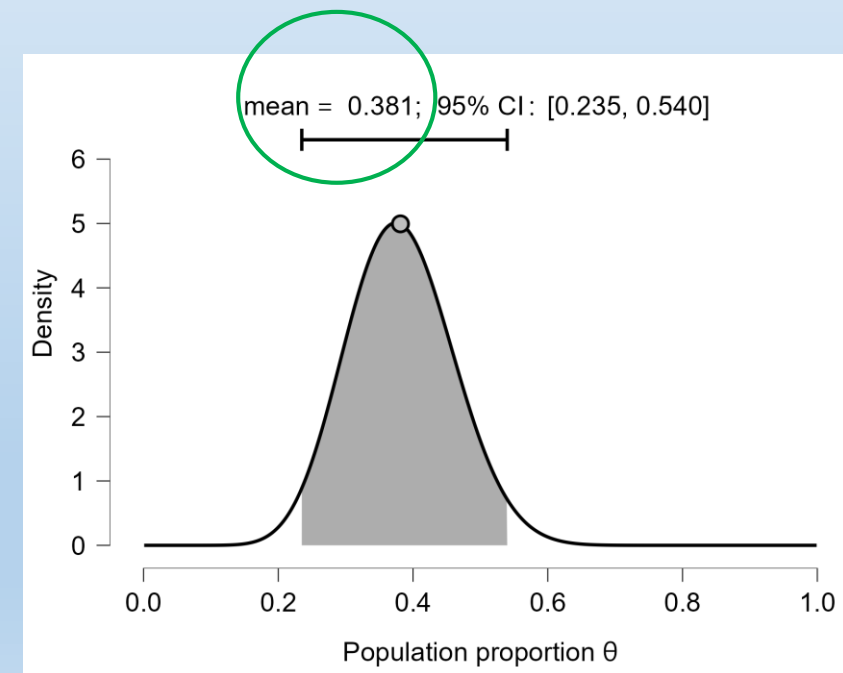
$$X \sim \text{Bin}(27, p)$$

Il nostro «**prior**» su p segue una distribuzione di probabilità «Beta» di parametri (3.3, 7.2):



L'intervallo di credibilità al 95% è : **[0.235 – 0.54]**

$$P(p|\text{data}) \cong P(\text{data}|p)P(p)$$



Relazione tra la **numerosità campionaria** e l'ampiezza dell'IC...

L'ampiezza totale dell'intervallo di confidenza per la media di una popolazione è data da:

$$d = 2 * z_{\frac{\alpha}{2}} * \sqrt{\frac{\sigma^2}{n}}$$


Costante della gaussiana (es: 1,96)

A parità del livello **1- α** scelto e della varianza l'ampiezza dell'intervallo dipende dalla dimensione campionaria n , al crescere della quale l'ampiezza si riduce.

In molti casi applicativi, la dimensione campionaria n è fissata in partenza e dipende dal budget a disposizione per l'estrazione del campione.

In altri casi (ad esempio negli studi clinico-epidemiologici) è molto importante fissare l'ampiezza d^* che l'intervallo **non può superare** e determinare la **dimensione campionaria minima** n^* che garantisce tale requisito, cioè tale per cui quando $n < n^*$ si ottiene un intervallo con ampiezza $d > d^*$ (ovviamente per tutti gli $n > n^*$ si ottiene un intervallo con ampiezza $d < d^*$).

Per effettuare il calcolo di n^* è sufficiente osservare che se deve essere:

$$2 * z_{\frac{\alpha}{2}} * \sqrt{\frac{\sigma^2}{n}} \leq d^*$$

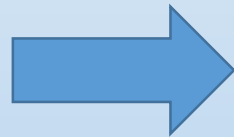
$$\sqrt{\frac{\sigma^2}{n}} \leq \frac{d^*}{2 * z_{\frac{\alpha}{2}}} \quad \Rightarrow \quad \frac{\sigma^2}{n} \leq \left(\frac{d^*}{2 * z_{\frac{\alpha}{2}}} \right)^2 \quad \Rightarrow \quad n \geq \left(\frac{2 * z_{\frac{\alpha}{2}} * \sigma}{d^*} \right)^2$$

Esempio: Da informazioni derivanti da una precedente analisi, si sa che la durata delle telefonate che arrivano al CUP di un ospedale si distribuisce in modo approssimativamente normale con media μ incognita e deviazione standard pari a 4 minuti.

Si desidera calcolare la dimensione campionaria minima necessaria per costruire un intervallo di confidenza per la stima della durata media delle chiamate al livello 95% che abbia un'ampiezza massima di 5 minuti.

La dimensione richiesta è data da:

$$n^* = \left(\frac{2 * z_{\alpha/2} * \sigma}{d^*} \right)^2 = \left(\frac{2 * z_{\alpha/2} * 4}{5} \right)^2 = 9,8$$



Bisogna avere un campione di almeno 10 telefonate

[si può in questi casi correggere la formula a posteriori utilizzando la costante della t di student poiché siamo nel caso di un piccolo campione, quindi sostituire 1.96 con il valore della t di Student con 9 gradi di libertà, ovvero 2,26 e ottenere un campione di almeno 13 telefonate]

La conoscenza di σ^2 è cruciale per la determinazione della dimensione campionaria ottimale.
Tale conoscenza può derivare da studi precedenti o da studi pilota.

E' comunque consigliabile sovrastimare la varianza (piuttosto che sottostimarla) in quanto è meglio utilizzare una dimensione campionaria *troppo elevata* che una *troppo bassa*.

Il calcolo della dimensione campionaria ottimale può essere fatto anche quando l'intervallo di confidenza è calcolato per una proporzione incognita p .

Naturalmente, in questo caso la precisione dello stimatore (e quindi l'ampiezza dell'intervallo) dipende dal valore assunto da p , che è incognito.

In particolare, ci si accorge che la funzione $p^*(1 - p)$ è una funzione concava che vale 0 quando $p = 0$ oppure 1 e raggiunge il suo massimo quando $p = 0.5$. Se ne deduce che a parità di dimensione campionaria e di livello di confidenza otterremo intervalli di confidenza generalmente più stretti quando p si trova vicino agli estremi 0 e 1, e più larghi quando p si trova in un intorno di 0.5.

E' dunque appropriato usare come misura cautelativa proprio $p=0.5$: $p^*(1-p)=0.25$

Se dunque desideriamo calcolare la dimensione minima richiesta per avere un intervallo per p che non superi l'ampiezza globale massima d^* dobbiamo cercare il minimo valore di n tale che:

$$2 * z_{\frac{\alpha}{2}} * \sqrt{\frac{0.25}{n}} \leq d^*$$

Dopo qualche passaggio si ottiene:

$$n^* = 0.25 * \left(\frac{2 * z_{\frac{\alpha}{2}}}{d^*} \right)^2$$

Se programmiamo uno studio epidemiologico per stimare la proporzione di soggetti con una certa malattia e desideriamo che l'intervallo di confidenza al livello $1 - \alpha = 0.95$ non superi l'ampiezza di ± 2 (4) , avremo bisogno di un minimo di:

$$n^* = 0.25 * \left(\frac{2 * 1.96}{0.04} \right)^2 = 2401$$

Dovremo estrarre dalla popolazione target un campione di almeno 2401 persone.

Sfortunatamente, non esiste una formula semplice e universale come quella utilizzata nell'inferenza frequentista per il calcolo della **dimensione del campione** quando si utilizza l'approccio bayesiano.

Il motivo principale è che l'ampiezza dell'intervallo di credibilità dipende non solo dalla dimensione del campione, ma anche da:

- **Distribuzione a priori:** La scelta della distribuzione a priori influisce notevolmente sulla forma della distribuzione a posteriori e, di conseguenza, sulla larghezza dell'intervallo di credibilità.
- **Definizione di intervallo di credibilità:** Esistono diversi modi per definire un intervallo di credibilità (es. HPD, quantili...).

*Possibile approccio: **simulare** ripetutamente la distribuzione a posteriori per diverse dimensioni del campione.*

Per ogni dimensione del campione, calcolare l'ampiezza dell'intervallo di credibilità.

Si sceglie la dimensione del campione che produce una ampiezza accettabile...
