Cap. 7 Distribuzioni campionarie

Popolazione e Campione

Una **popolazione** è l'insieme di tutte le unità oggetto di studio

- Tutti i potenziali votanti nelle prossime elezioni
- Tutti i pezzi prodotti oggi
- Tutti gli scontrini di novembre

Un **campione casuale** è un sottoinsieme della popolazione scelto in modo che
sia nota a probabilità di estrarre ogni unità

- Alcuni votanti selezionati casualmente per un'intervista
- Alcuni pezzi selezionati per un test di distruzione
- Alcuni scontrini selezionati casualmente per una verifica

Inferenza statistica

Come si può risalire alla descrizione della popolazione disponendo solo delle informazioni estratte dal campione?

Se il campione è casuale si può fare una **stima** di certe caratteristiche della popolazione e si può fornire una indicazione dell'**errore di campionamento**

La magia è possibile solo se i dati sono raccolti in modo opportuno. Per esempio con un metodo di campionamento casuale semplice con o senza ripetizione

Esempio (Sondaggi)

In una popolazione di 100 milioni di votanti per il Presidente USA ci sono il 40% di favorevoli a Hilary Clinton.

Quindi se scegliamo casualmente un votante (in modo che ogni votante abbia la stessa probabilità di essere estratto)
la probabilità di successo è **$p = 0.4$**

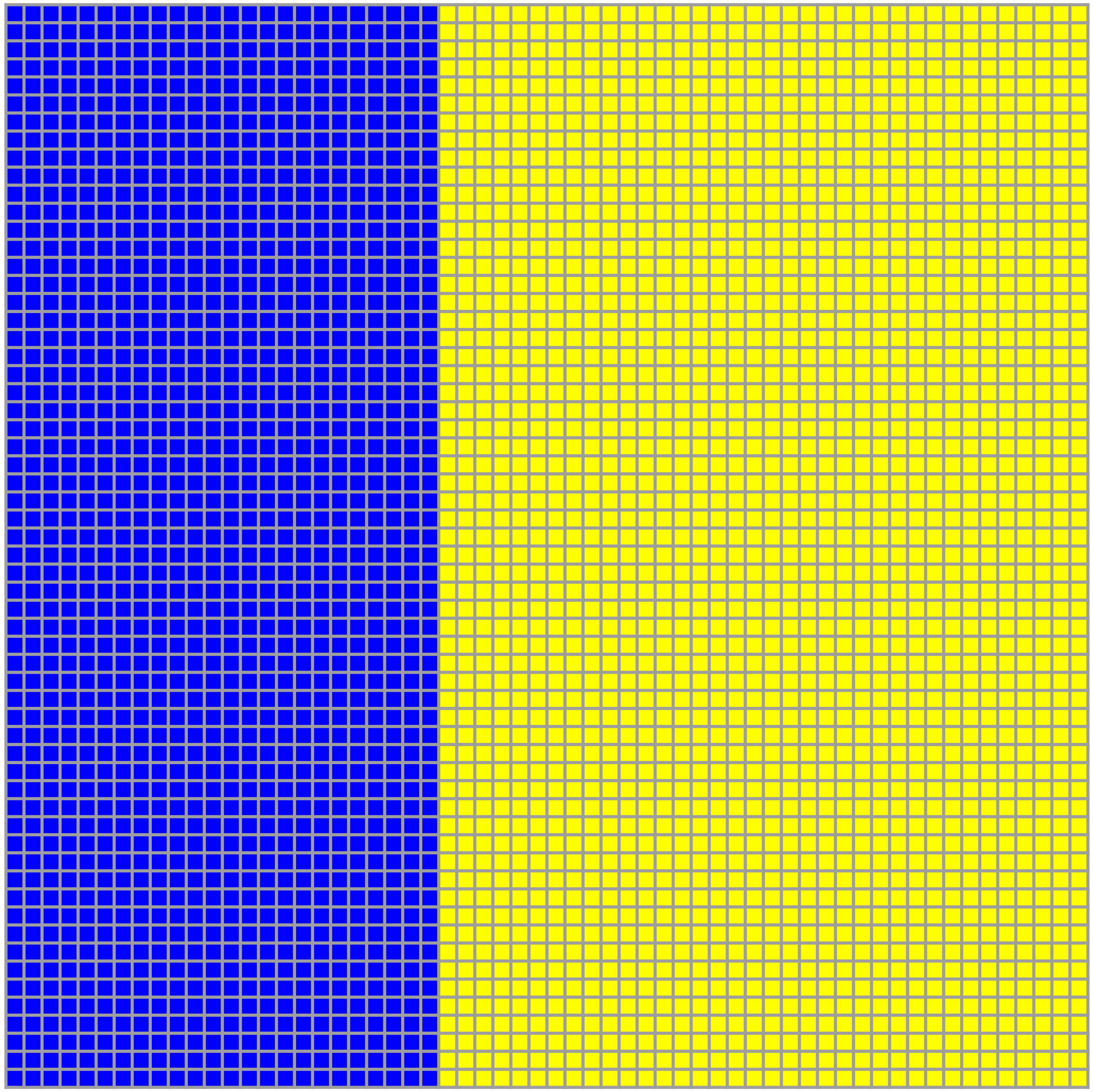
Supponiamo di estrarre un campione casuale con ripetizione di **$n = 200$** votanti e di **non sapere il valore di p**

1) Possiamo stimare p da questi dati?

2) Quant'è l'errore che si commette usando solo 200 votanti?

Perchè funziona con campioni casuali

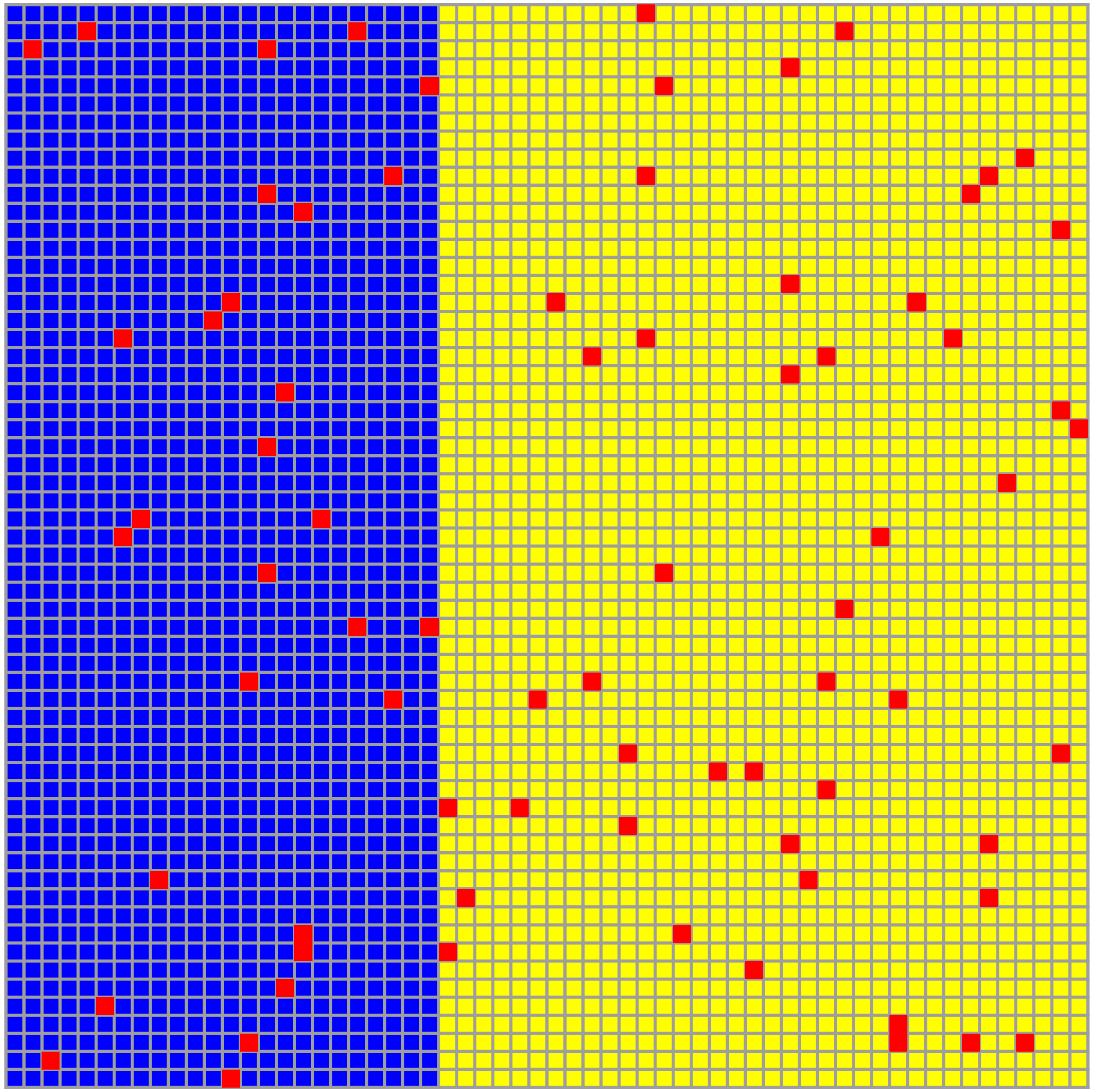
Clinton



Resto

Perchè funziona con campioni casuali

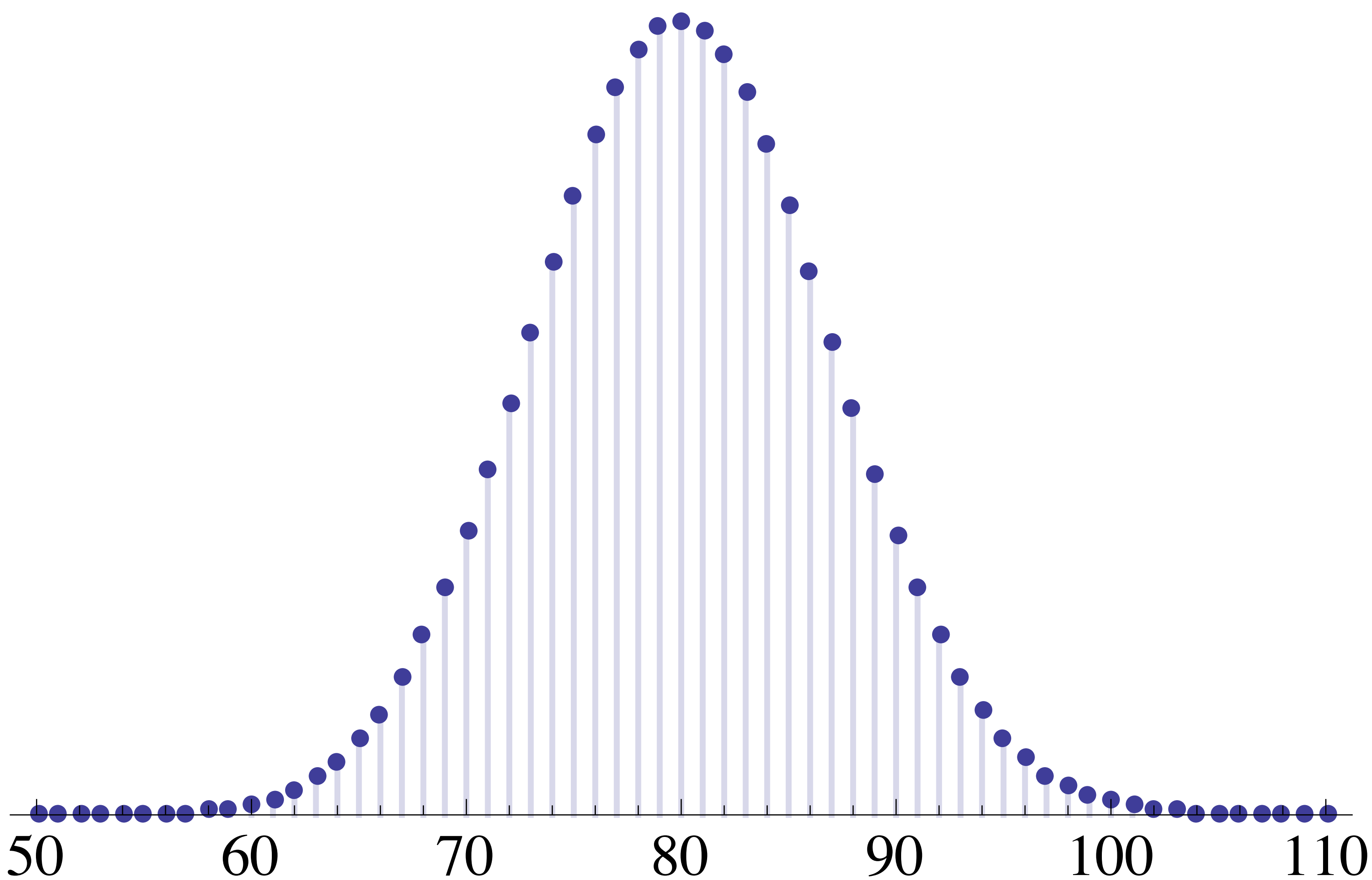
Clinton



Resto

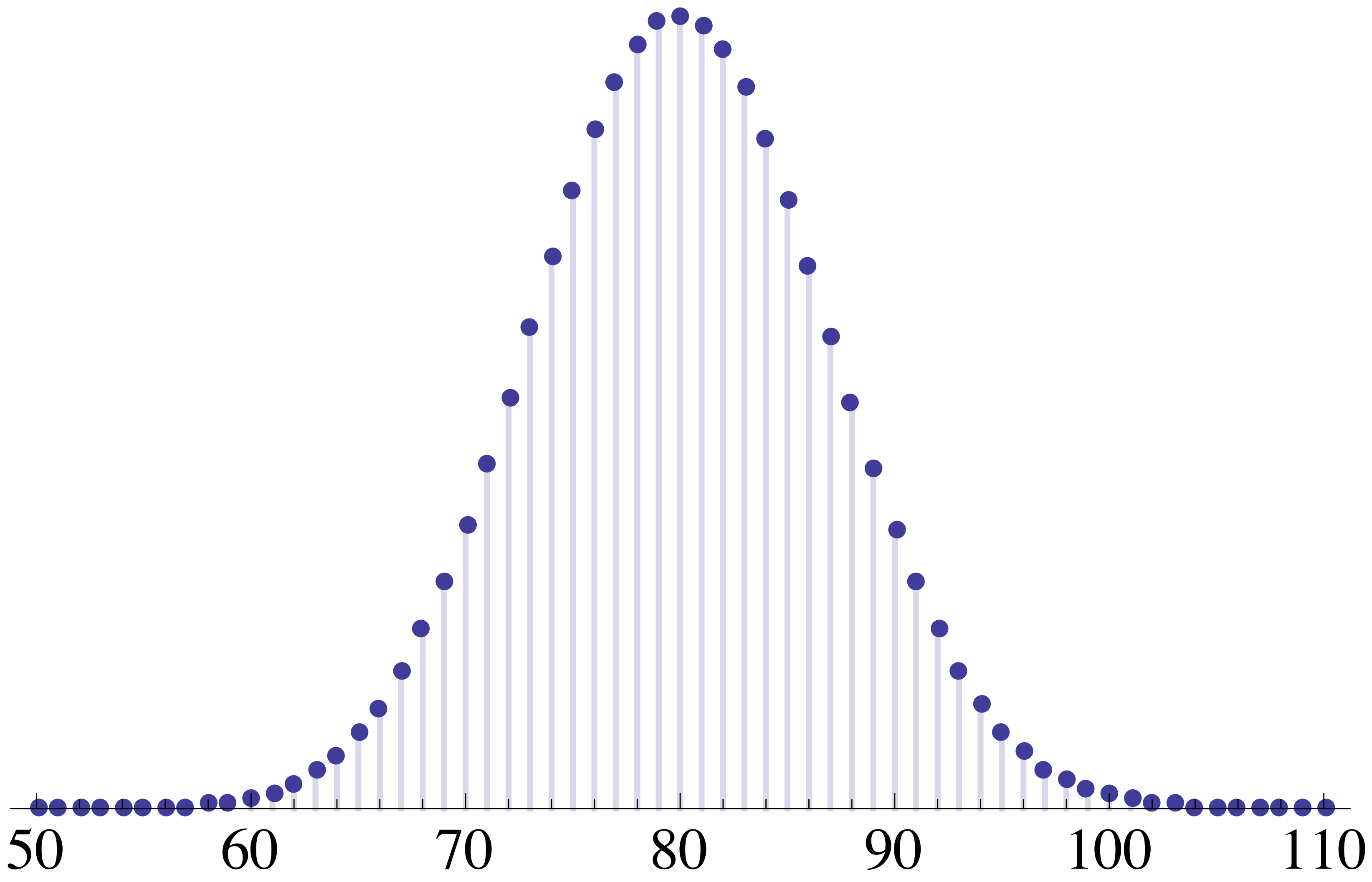
Distribuzione del numero di successi

In un campione con ripetizione di 200 individui
il numero di successi X (voti a favore di Clinton)
ha **distribuzione Binomiale** (p = 0.4, n = 200)



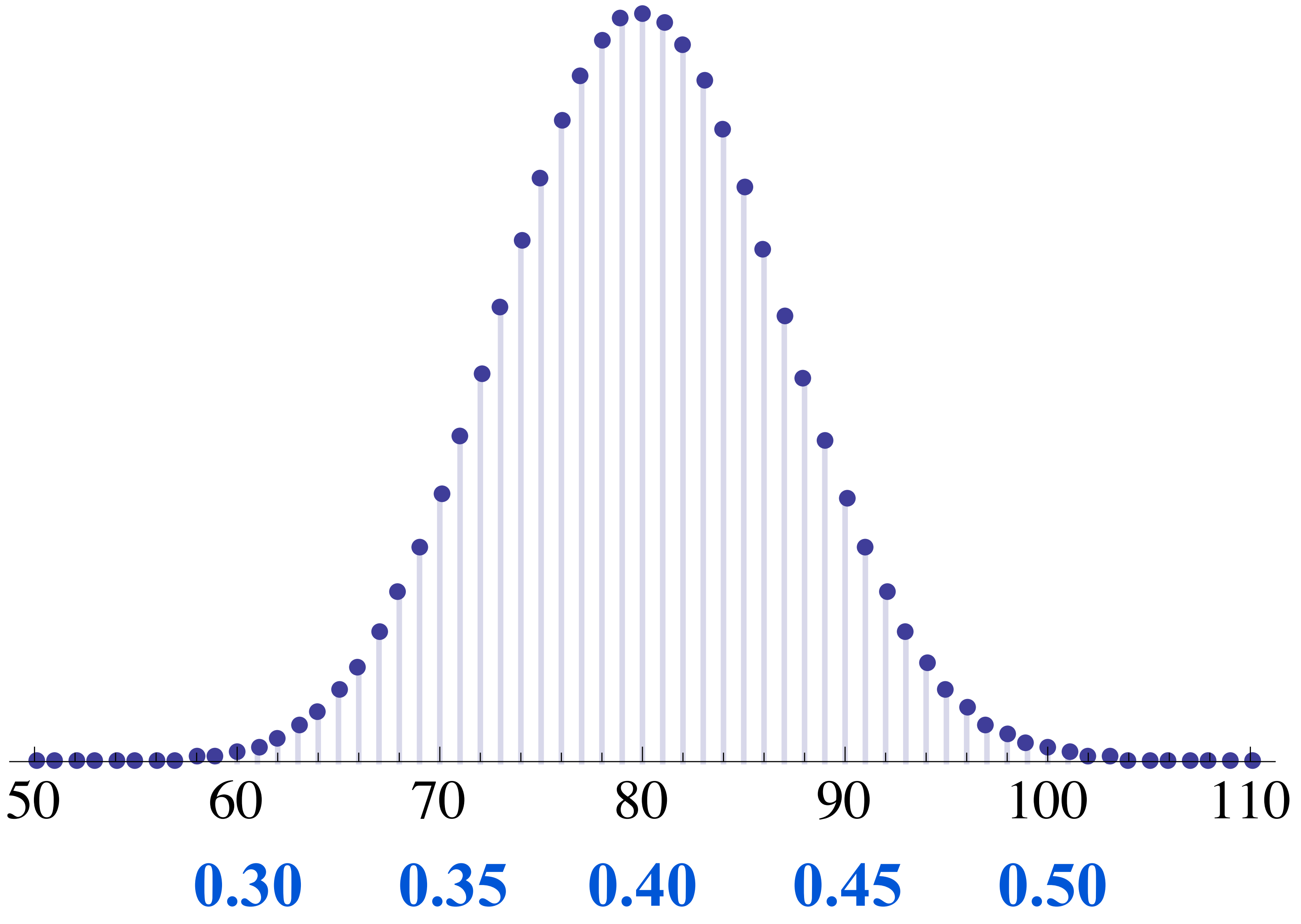
Distribuzione campionaria

Questa distribuzione Binomiale (p = 0.4, n = 200) si chiama
distribuzione campionaria del numero di successi



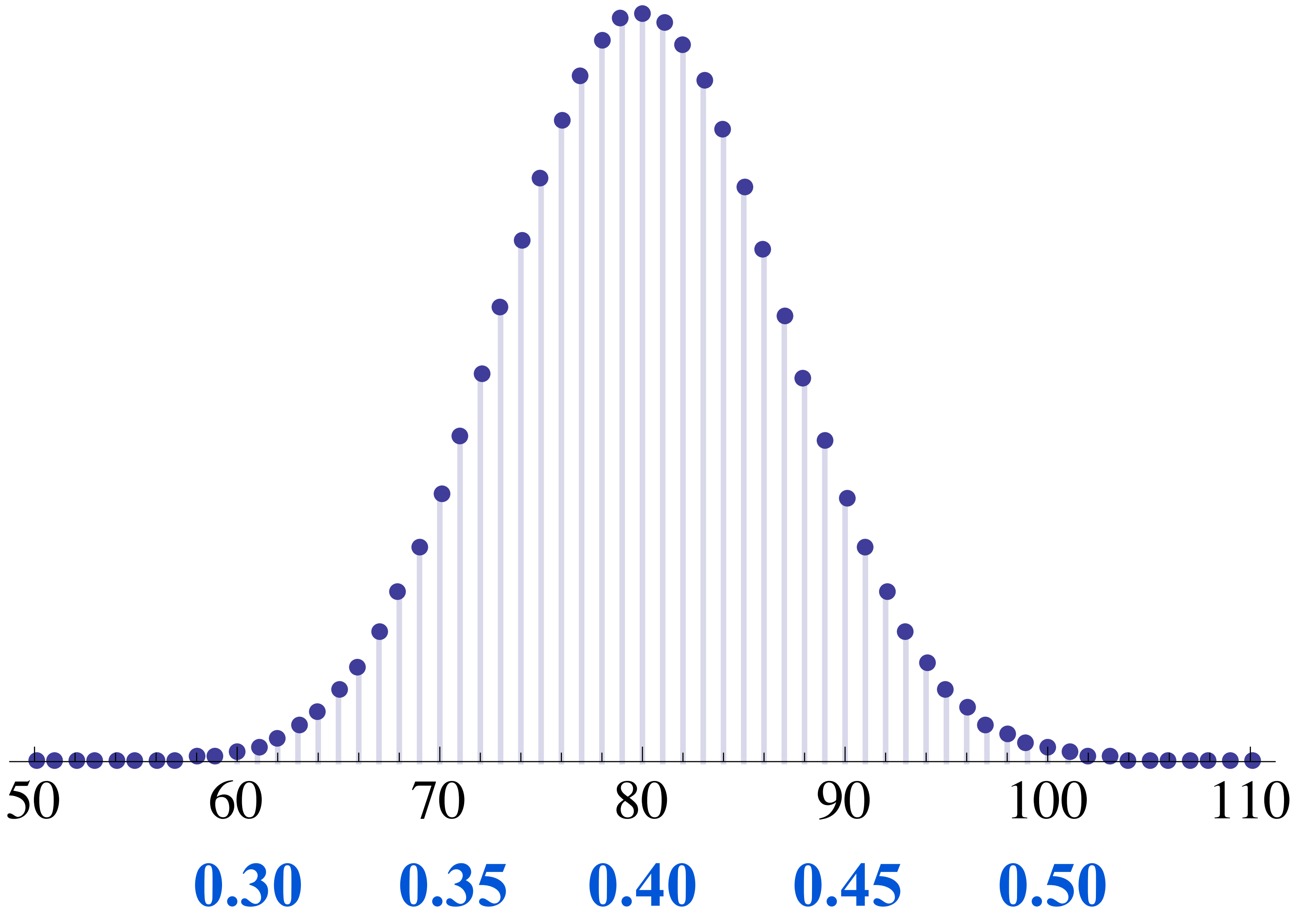
Distribuzione campionaria della proporzione

La **distribuzione campionaria della proporzione di successi**



Distribuzione campionaria della proporzione

La proporzione di successi è S / n = #successi / n



Media = 0.4

Distribuzione campionaria della proporzione

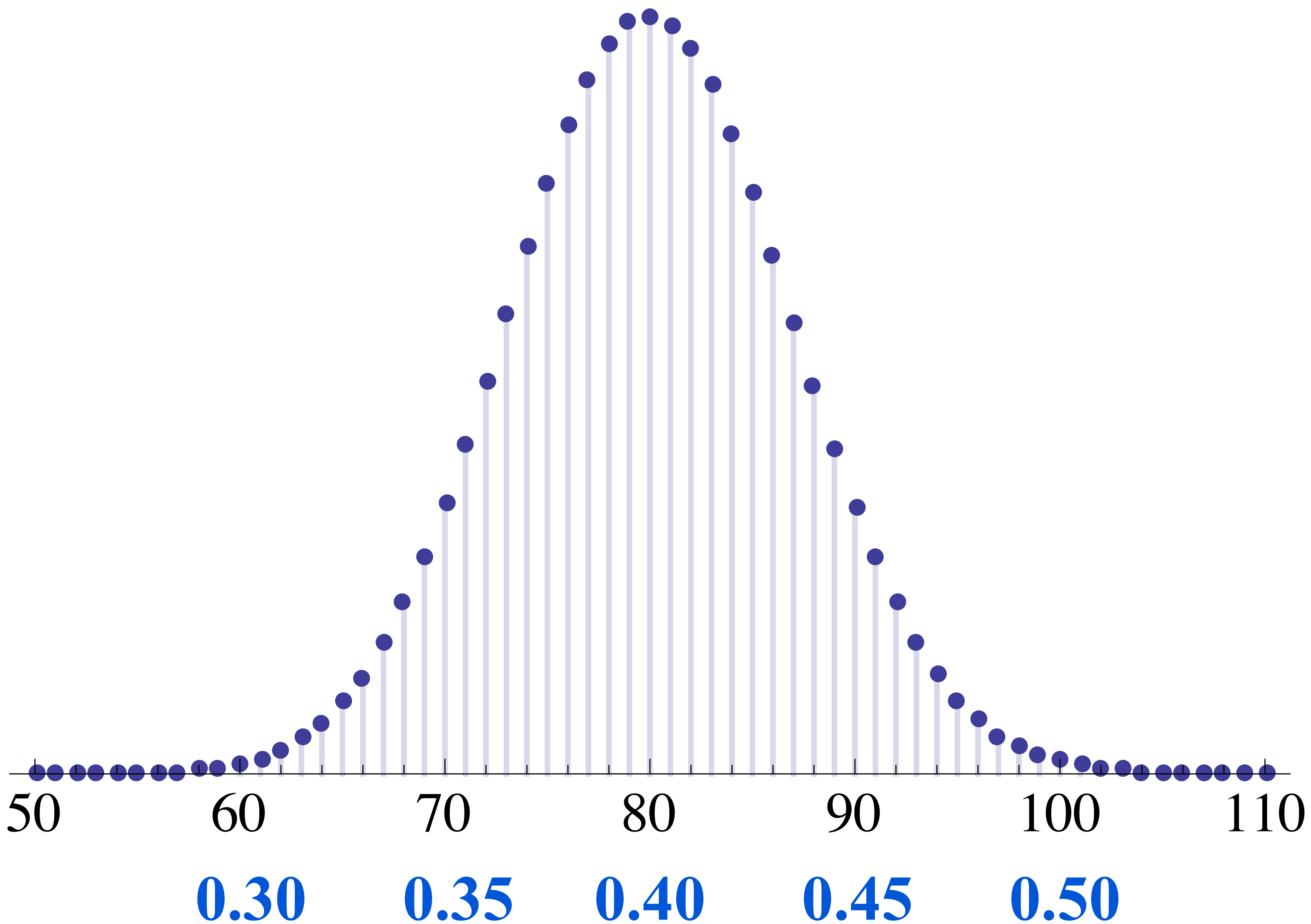
La proporzione di successi è S / n = #successi / n

Se il campione è con ripetizione ha distribuzione Binomiale
con media

$$E(S/n) = E(S)/n = np/n = p$$

**Quindi, se nel campione calcolo la proporzione
di voti per Clinton mi aspetto che sia
proprio intorno alla media p = 0.4**

Distribuzione campionaria della proporzione



Inoltre, per la regola empirica mi aspetto di trovare la proporzione di voti a Obama compresa tra 0.3 e 0.5 nel 99% dei casi

Regola empirica applicata alla proporzione

$$\mathrm{var}(X/n) = \frac{1}{n^2} npq = \frac{pq}{n}$$

$$\sigma(X/n) = \sqrt{\frac{pq}{n}} = 0.035$$

La deviazione standard è 3.5%

Quindi nel 99% dei casi troveremo la proporzione X/n compresa nell'intervallo

$$[p - 3\sqrt{\frac{pq}{n}} \quad p + 3\sqrt{\frac{pq}{n}}]$$

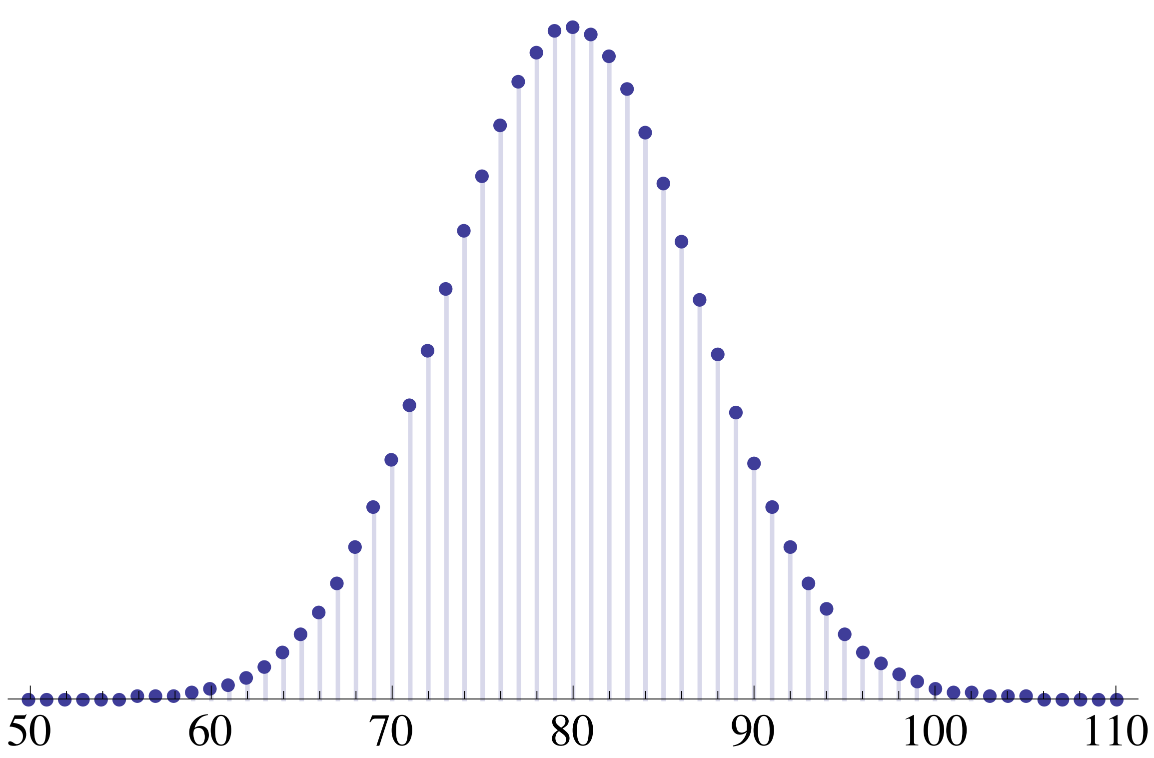
3 sigma sono circa il 10%

0.4 - 3 * 0.035 = **0.3**   0.4 + 3 * 0.035 = **0.5**

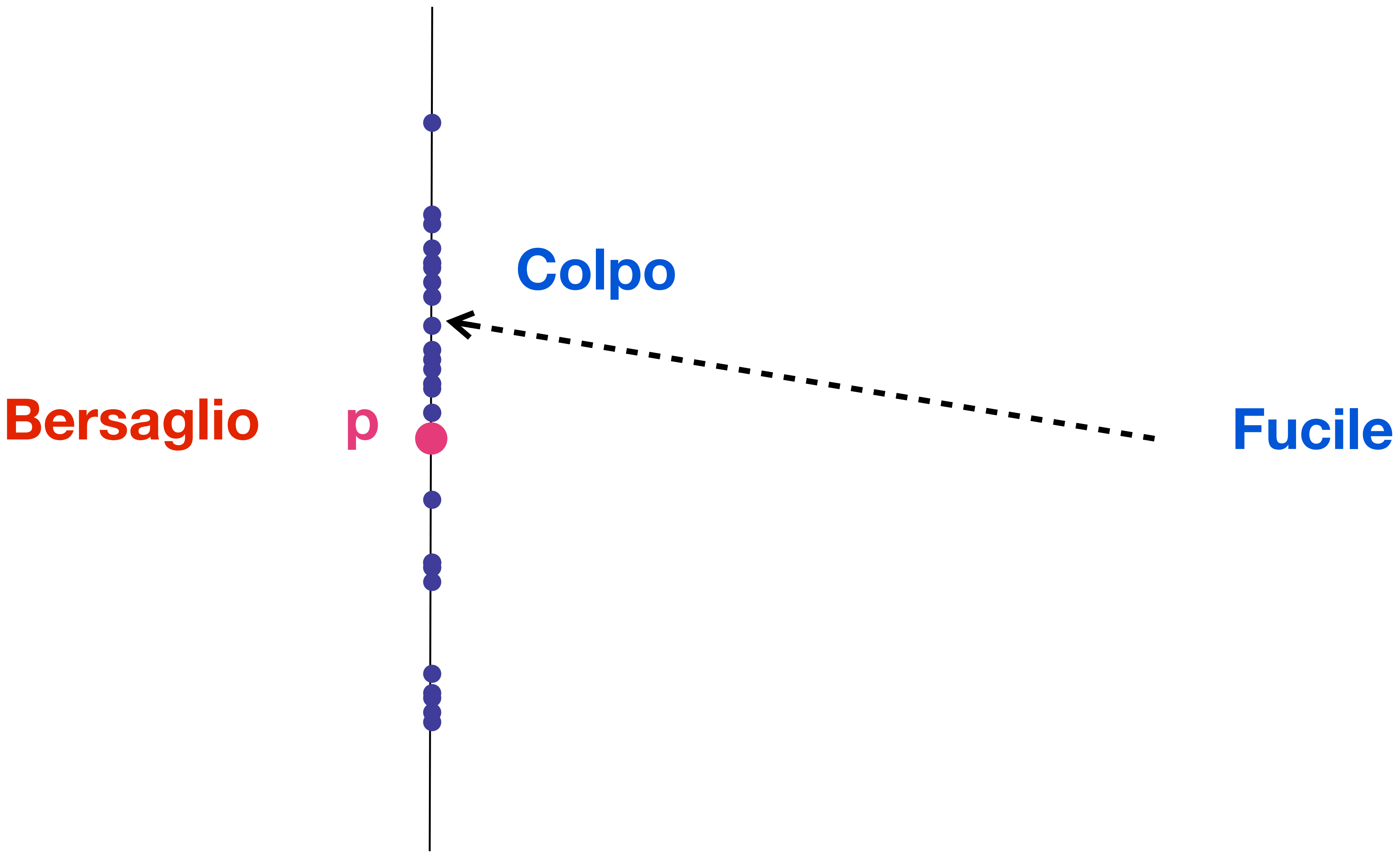
Campionamento ripetuto

La distribuzione campionaria binomiale
si può interpretare come la distribuzione
della proporzione di successi **nel campionamento ripetuto**

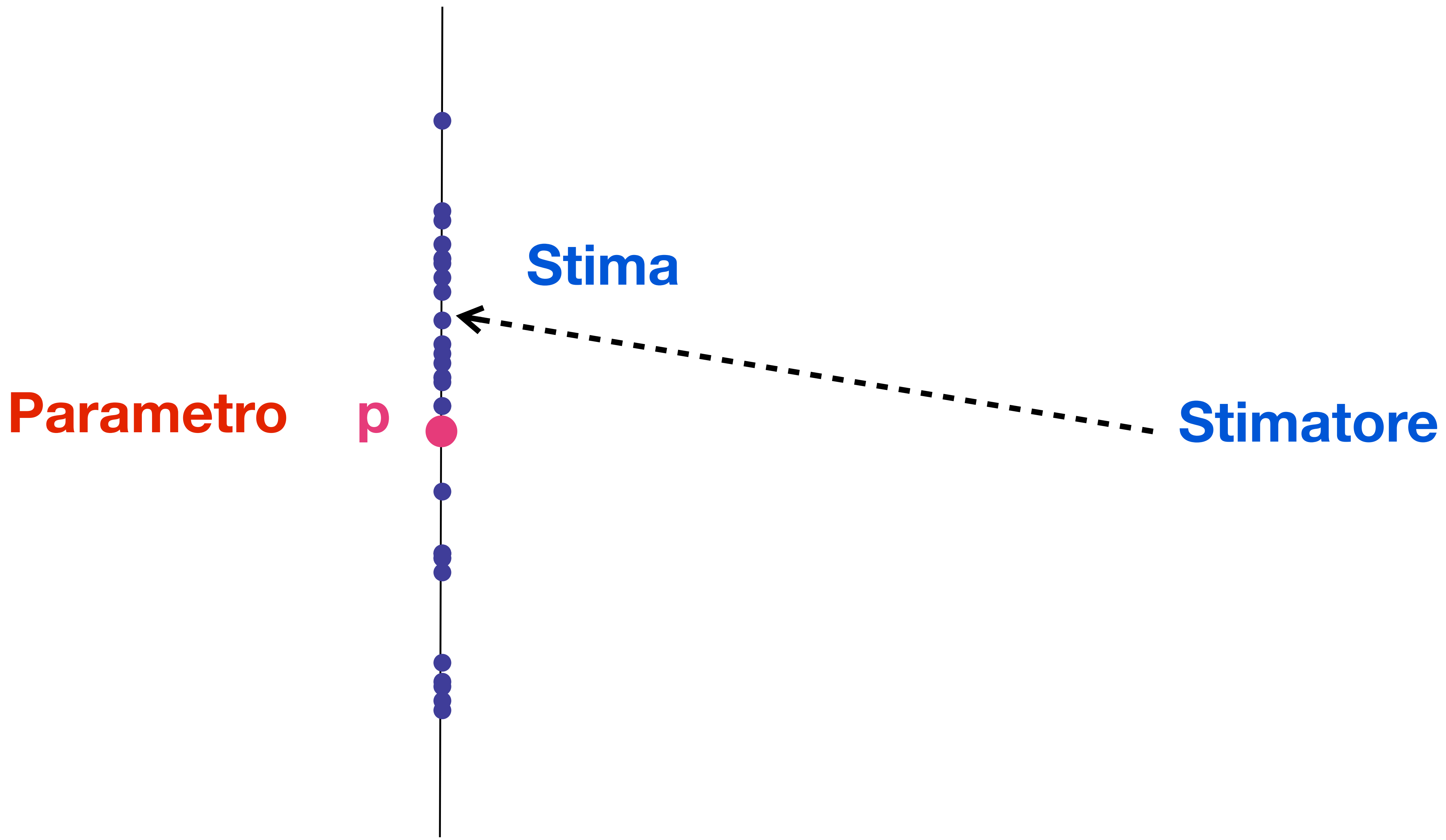
Cioè è la distribuzione della proporzione **nel lungo andare**,
immaginando di continuare ad estrarre campioni all'infinito



Analogia

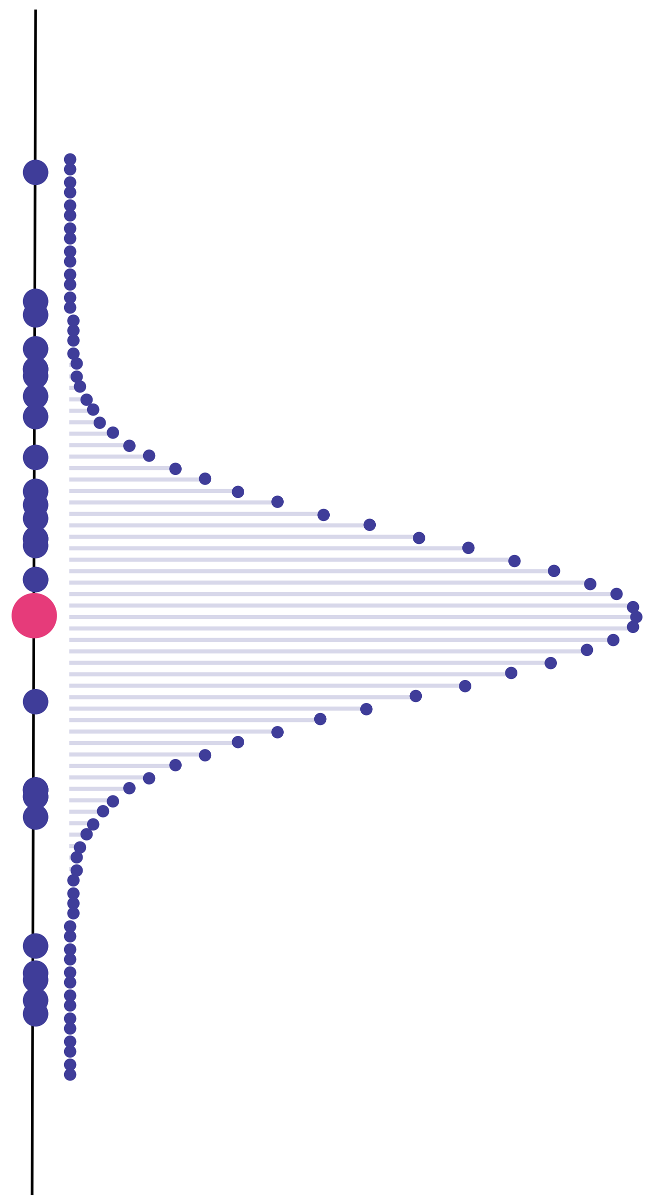


Analogia



Analogia

Parametro **p**



Distribuzione campionaria

È la rosa di colpi del fucile

Campionamento in generale

Singola osservazione di 1 unità dalla popolazione
È la variabile aleatoria X che descrive la popolazione

Campione casuale con ripetizione di n unità
È formato da *n* variabili aleatorie $X_1, X_2, \ldots, X_n$

1) *indipendenti*
2) e con distribuzione identica a quella di X

Campionamento (segue)

L'insieme di tutti i possibili campioni di dimensione n
si descrive con un un'n-upla di **variabili aleatorie**

$$X_1, \ldots, X_n$$

Si dice talvolta
Universo dei campioni

Il campione osservato è invece un'n-upla di **numeri**

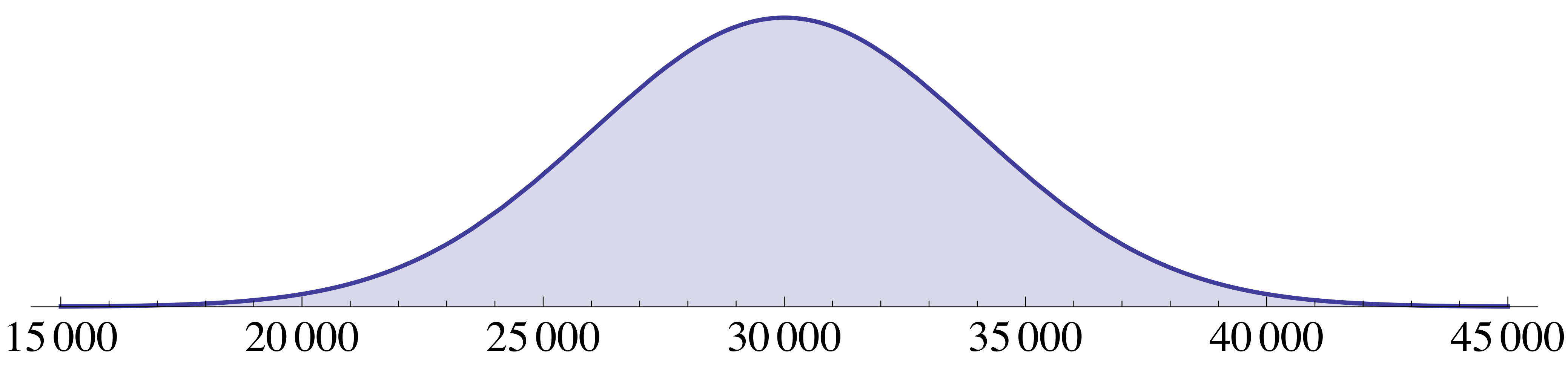
$$x_1, \ldots, x_n$$

Un altro esempio

Un produttore di pezzi di ricambio per auto dice che le sue candele hanno una **durata media di 30000** km con una **deviazione standard di 4000** km.

Dice inoltre che la durata X ha **distribuzione normale**

Si estrae un campione casuale di **16** candele

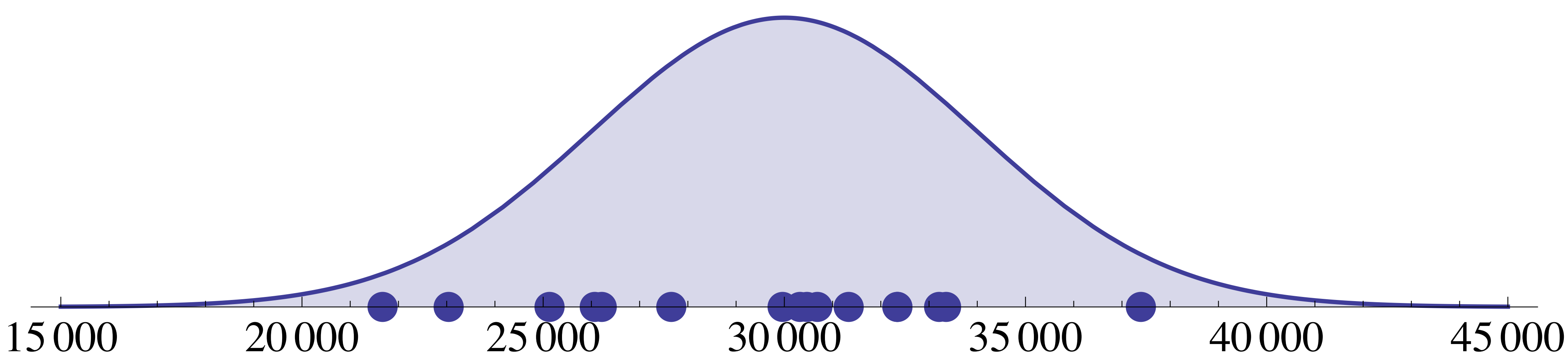


Un altro esempio

Un produttore di pezzi di ricambio per auto dice che le sue candele hanno una **durata media di 30000** km con una **deviazione standard di 4000** km.

Dice inoltre che la durata X ha **distribuzione normale**

33338, 31304, 27656, 29952, 32327, 26199, 30353, 30658, 25105, 23070, 32334, 26099, 30495, 33185, 37409, 21689



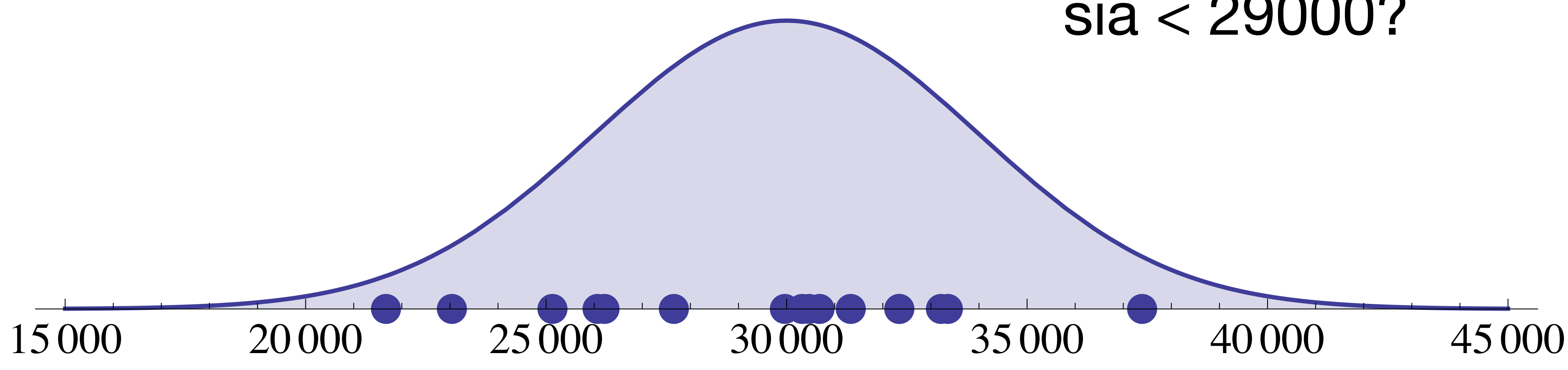
Un altro esempio

Un produttore di pezzi di ricambio per auto dice che le sue candele hanno una **durata media di 30000** km con una **deviazione standard di 4000** km.

Dice inoltre che la durata X ha **distribuzione normale**

Media = 29448.3

Qual è la probabilità che **la media campionaria** sia < 29000?



Che cos'è la media campionaria?

Un campione di durate di 16 candele è

$$X_1, X_2, \ldots, X_{16}$$

In cui le variabili aleatorie sono indipendenti e distribuite come N(30000, DS = 4000)

La media campionaria è allora la combinazione lineare

$$\bar{X} = (X_1 + X_2 + \cdots + X_{16})/16$$

Che distribuzione ha?

Distribuzione della media campionaria

$X_1, X_2, \ldots, X_{16}$

sono variabili aleatorie **indipendenti e normali** N(30000, DS = 4000)

Allora la media campionaria

$$\bar{X} = (X_1 + X_2 + \cdots + X_{16})/16$$

ha **distribuzione normale con media 30000 e deviazione standard 1000**

Stessa media, deviazione standard minore!

Perché?

1) La distribuzione è normale perché è combinazione lineare di variabili normali

2) La media è la stessa perché

$$
\begin{aligned}
E[\bar{X}] &= E[(X_1 + \cdots + X_{16})/16] \\
&= [E(X_1) + \cdots E(X_n)]/16 \\
&= [\mu + \cdots + \mu]/16 \\
&= \mu
\end{aligned}
$$

Perché?

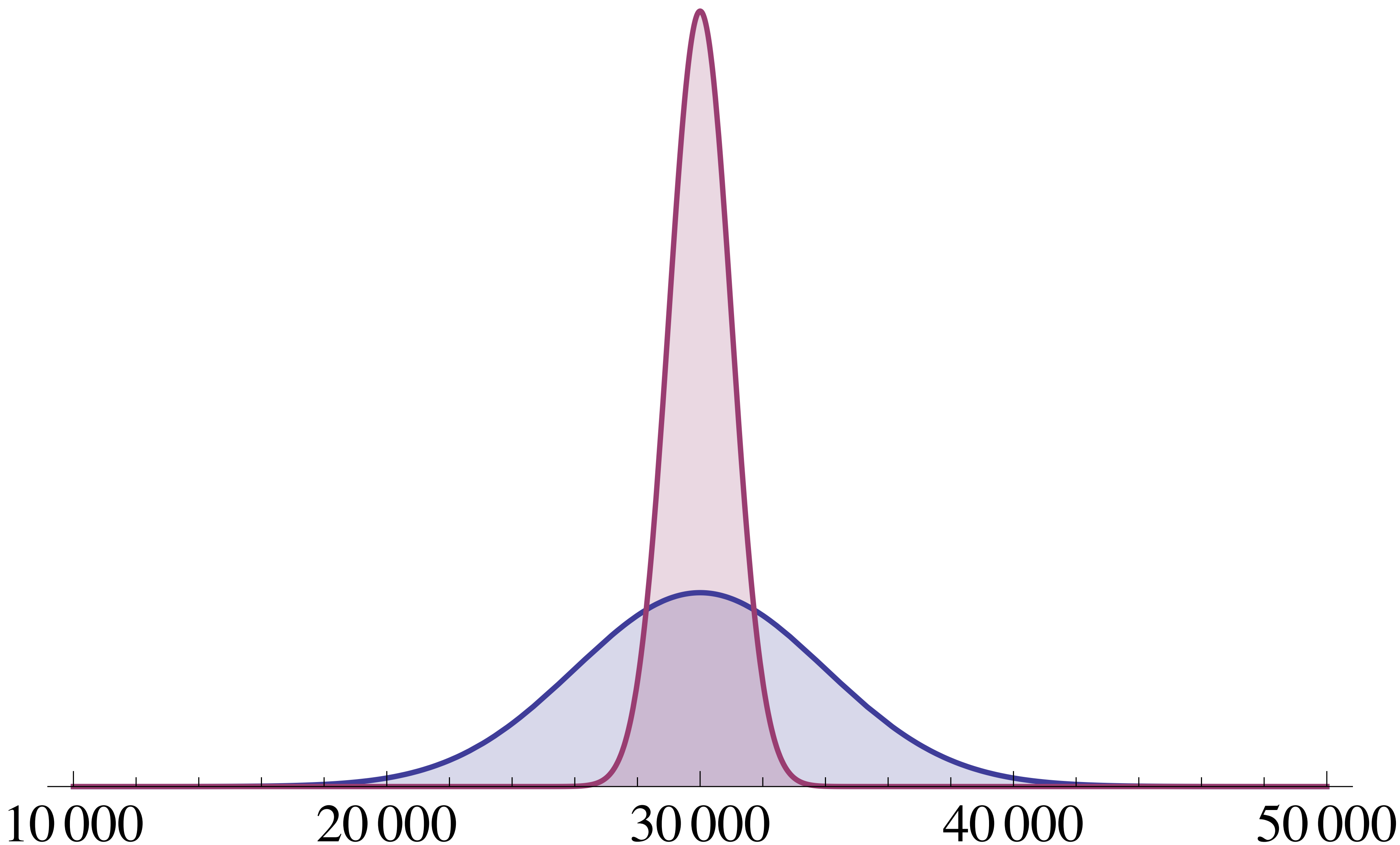
2) La varianza è minore perché

$$\begin{aligned}
\mathrm{var}[\bar{X}] &= \mathrm{var}[(X_1 + \cdots + X_{16})/16] \\
&= [\mathrm{var}(X_1) + \cdots \mathrm{var}(X_n)]/16^2 \\
&= [\sigma^2 + \cdots + \sigma^2]/16^2 \\
&= \sigma^2/16
\end{aligned}$$

Quindi $\sigma(\bar{X}) = \sigma/4$

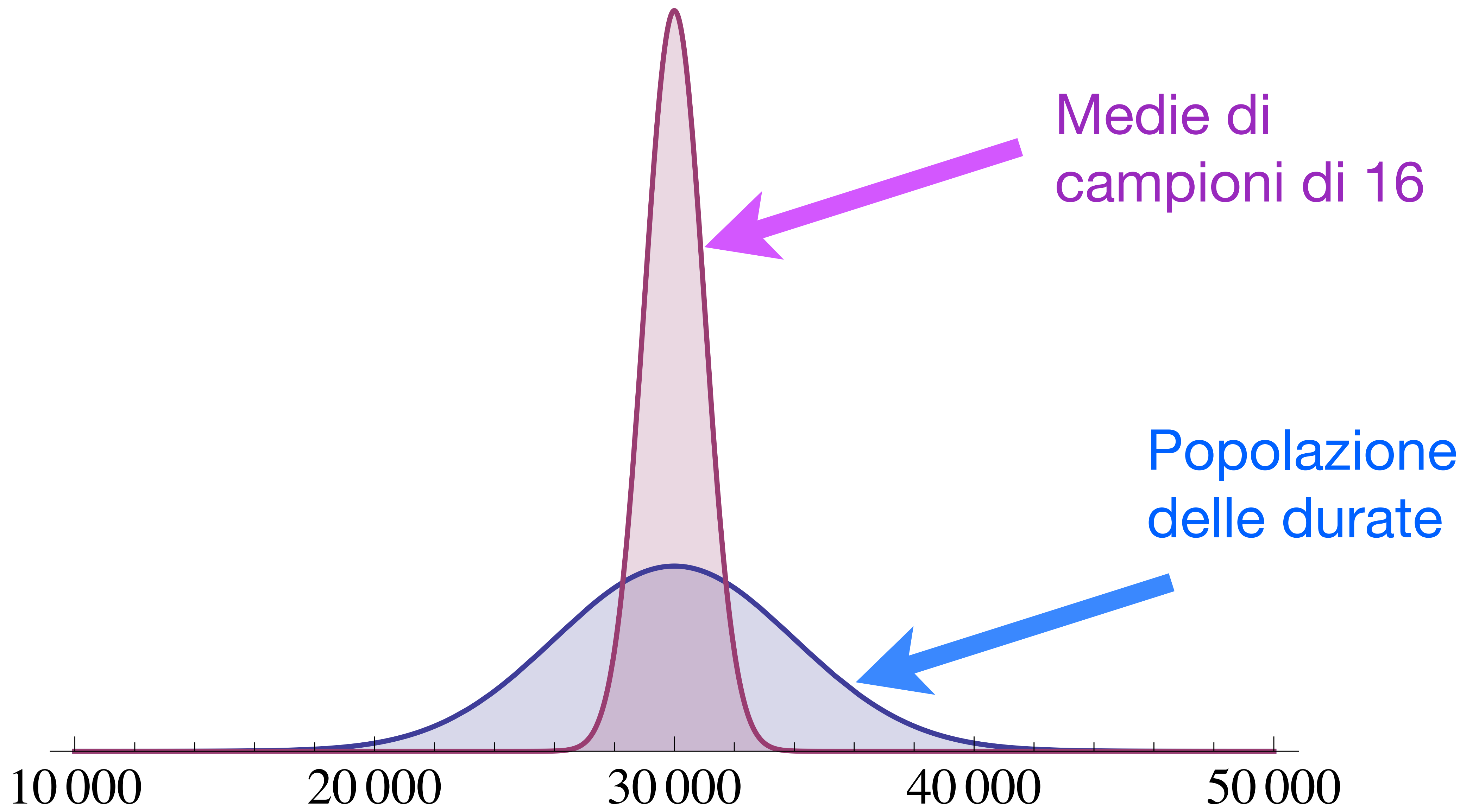
Conclusione

La media campionaria delle durate delle 16 candele
ha distribuzione N(30000, DS =1000)



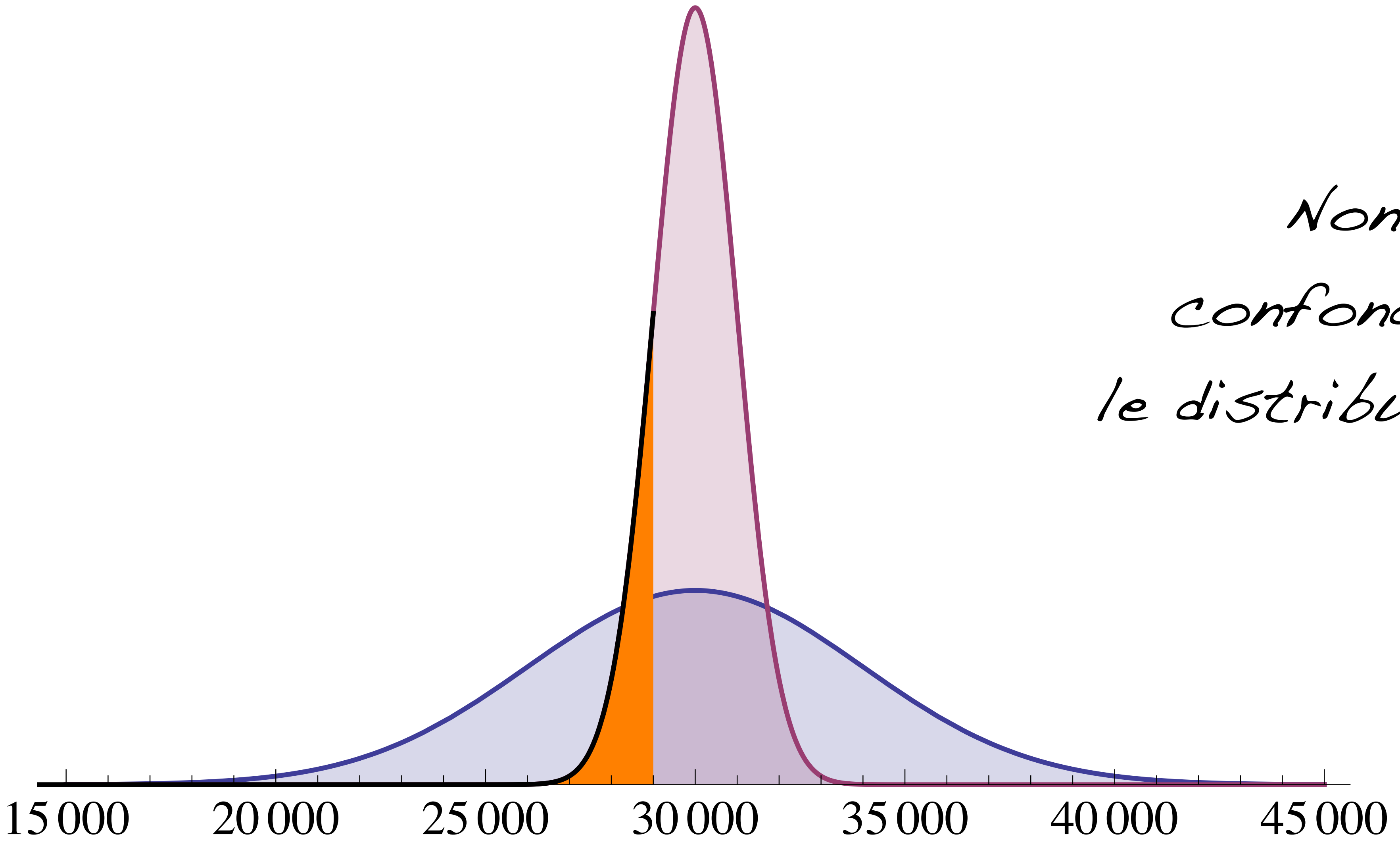
Conclusione

Le medie di campioni di dimensione 16 sono **molto più concentrate** attorno a 30000



Calcolo della probabilità $P(\bar{X} < 29000)$

$$
\begin{aligned}
P(\bar{X} < 29000) &= P(Z < (29000 - 30000)/1000) \\
&= P(Z < -1) = 1 - P(Z < 1) \\
&= 1 - 0.8413 = 0.1586
\end{aligned}
$$



Non confondere le distribuzioni!

Distribuzione campionaria della media

Dato un campione casuale da una popolazione con distribuzione di probabilità normale $X \sim N(\mu, \sigma^2)$

la sua media campionaria $\bar{X} = (X_1, + \cdots + X_n)/n$

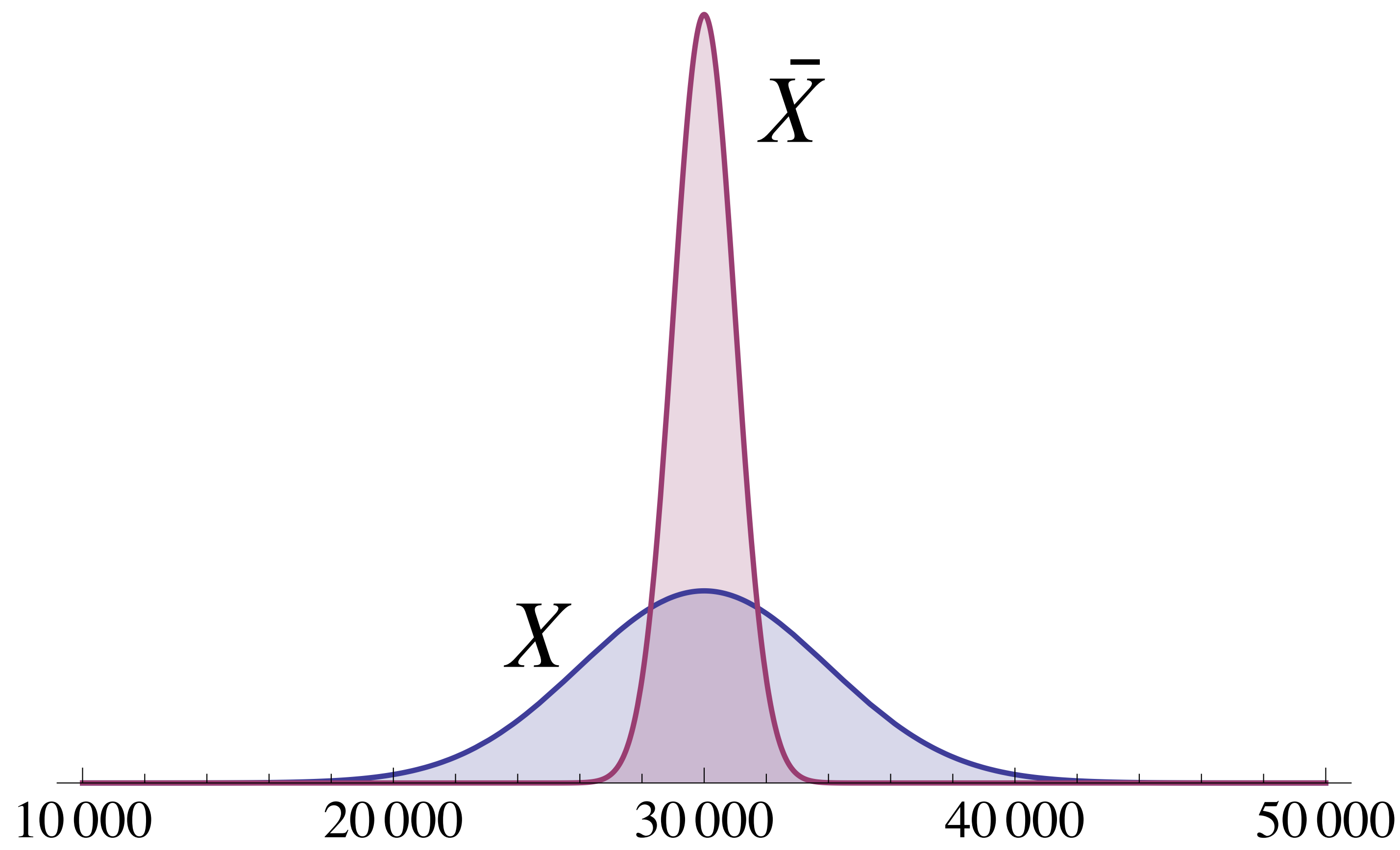
ha distribuzione campionaria normale

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

È centrata sulla media della popolazione

Ha una varianza più piccola che decresce all'aumentare di n

Distribuzione campionaria della media



$$\bar{X} \sim N(\mu, \sigma^2/n)$$

È centrata sulla
media della popolazione

Ha una varianza più piccola
che decresce all'aumentare di n

Stimatore non distorto

Non si può valutare se una specifica **stima** è buona o no

Si può valutare se lo **stimatore** è buono nel lungo andare

Uno stimatore si dice **corretto** o **non distorto** quando in media è centrato sul parametro da stimare

La media campionaria è uno stimatore corretto della media della popolazione perché

$$E(\bar{X}) = \mu$$

La proporzione di successi nel campione è uno stimatore corretto della probabilità p di successo perché

$$E(\hat{P}) = p$$

Variabilità dello stimatore

Uno stimatore corretto anche solo per campioni grandi è una buona cosa. Siamo infatti sicuri di non sovra- o sotto-stimare

Tuttavia è essenziale sapere qual è l'errore che si commette, cioè quanto si va vicini al bersaglio

L'errore che si commette è misurato dalla **deviazione standard della distribuzione campionaria dello stimatore**.

Questa si dice **errore standard dello stimatore**

Errore standard per la stima della media

La media campionaria si usa per stimare la media di una popolazione normale

La sua distribuzione di probabilità descrive come si comportano le stime nel campionamento ripetuto

La deviazione standard $\sigma_{\bar{X}} = \sqrt{\mathrm{var}(\bar{X})} = \sigma / \sqrt{n}$

indica quanto le stime sono variabili intorno al valore da stimare

Si chiama **errore standard** della stima.

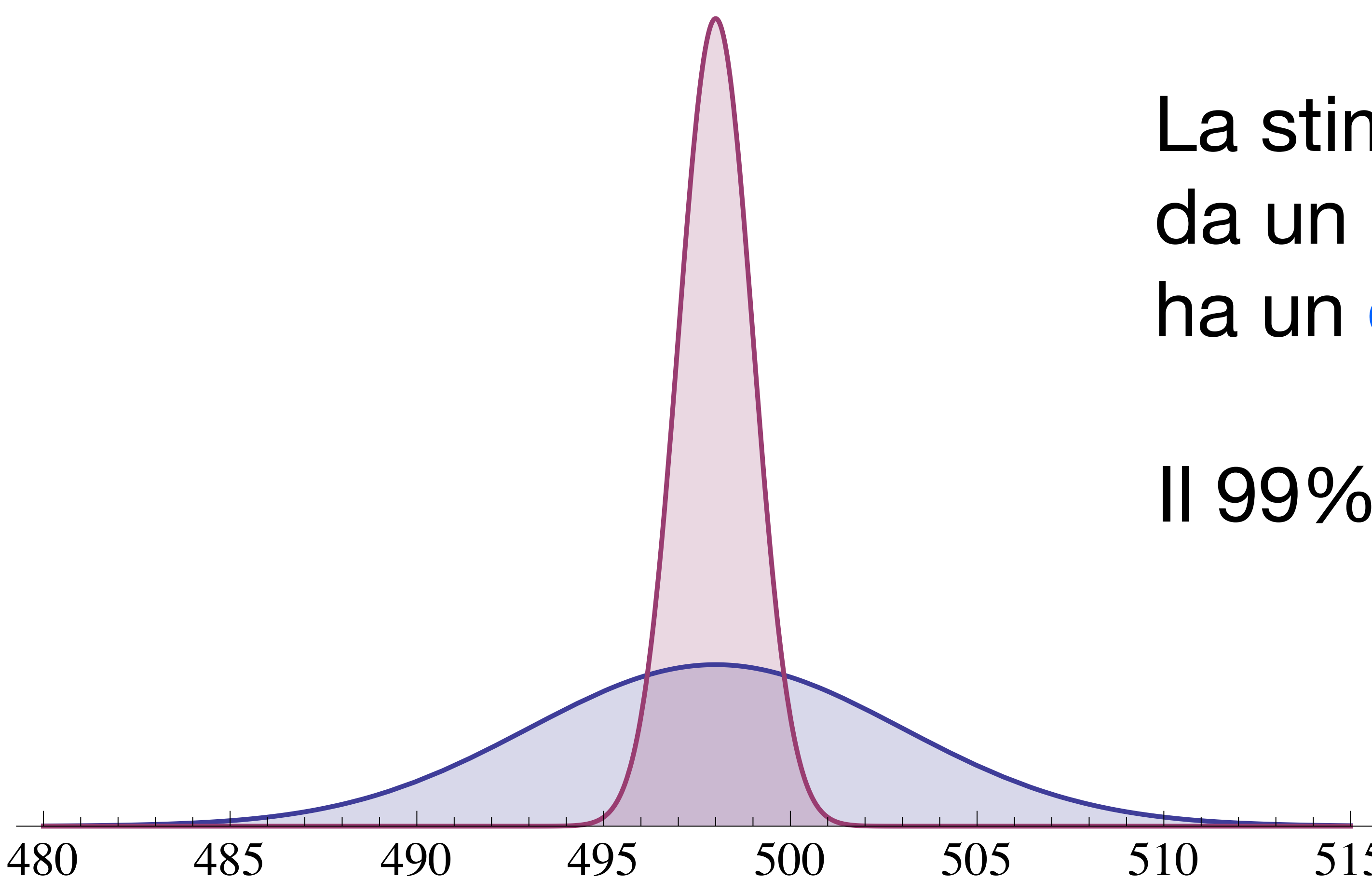
Esempio

Un produttore di pasta ha una macchina che riempe le scatole da mezzo kg. Si sa che le scatole riempite hanno un peso netto che si distribuisce **normalmente con media incognita e deviazione standard 5g**

Supponiamo che il produttore voglia stimare con un campione di 25 o 100 scatole quant'è la media di un pacco di pasta prodotto dalla sua macchina. Qual è l'errore standard della stima?

Esempio

Un produttore di pasta ha una macchina che riempe le scatole da mezzo kg. Si sa che le scatole riempite hanno un peso netto che si distribuisce **normalmente con media incognita e deviazione standard 5g**



vera media = 498g

La stima che ottengo da un campione di 25 casi ha un **errore standard di 5/5 = 1g**

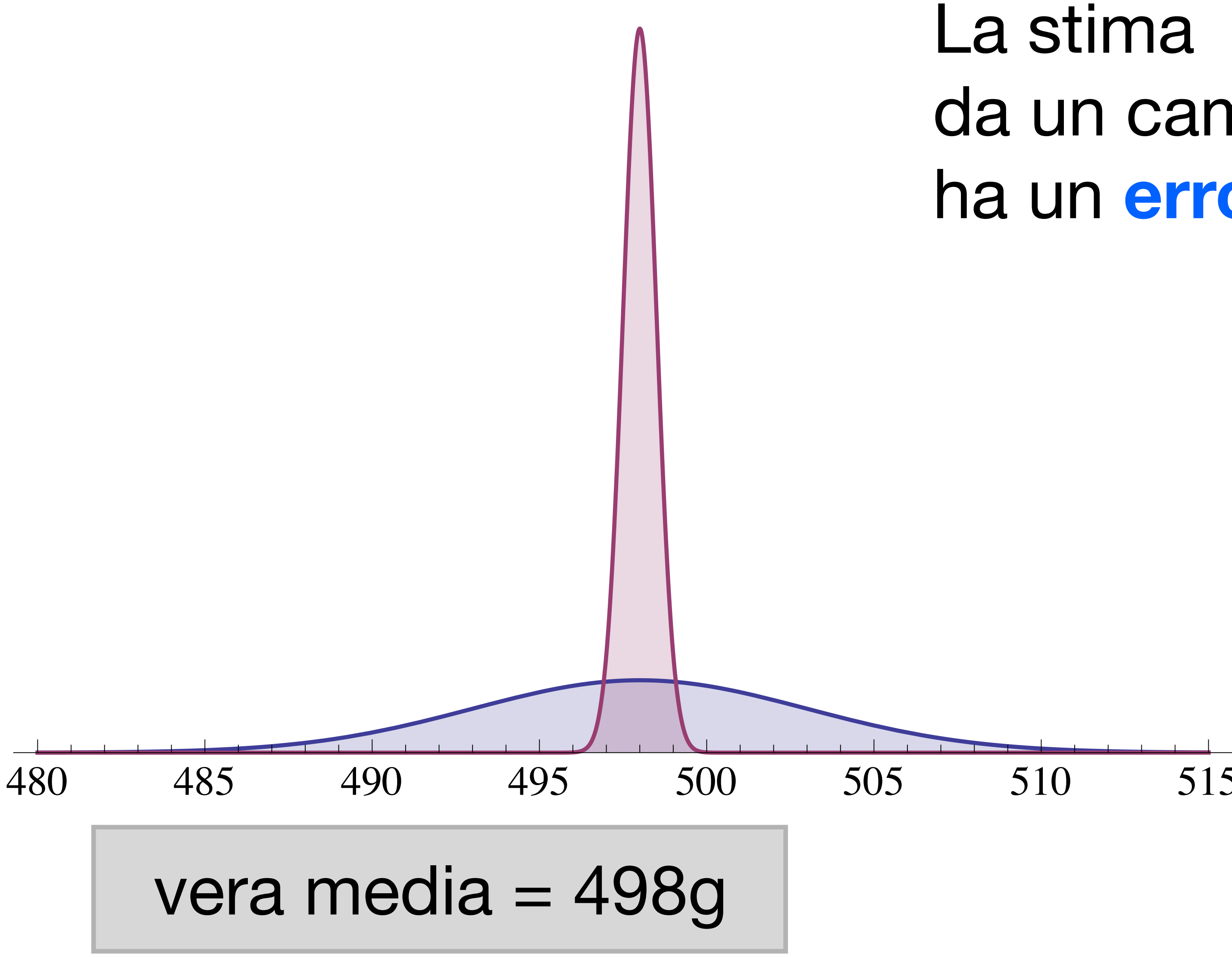
Il 99% delle stime cadrà in

$$[\mu - 3\text{g}, \quad \mu + 3\text{g}]$$

Esempio - aumento della dimensione

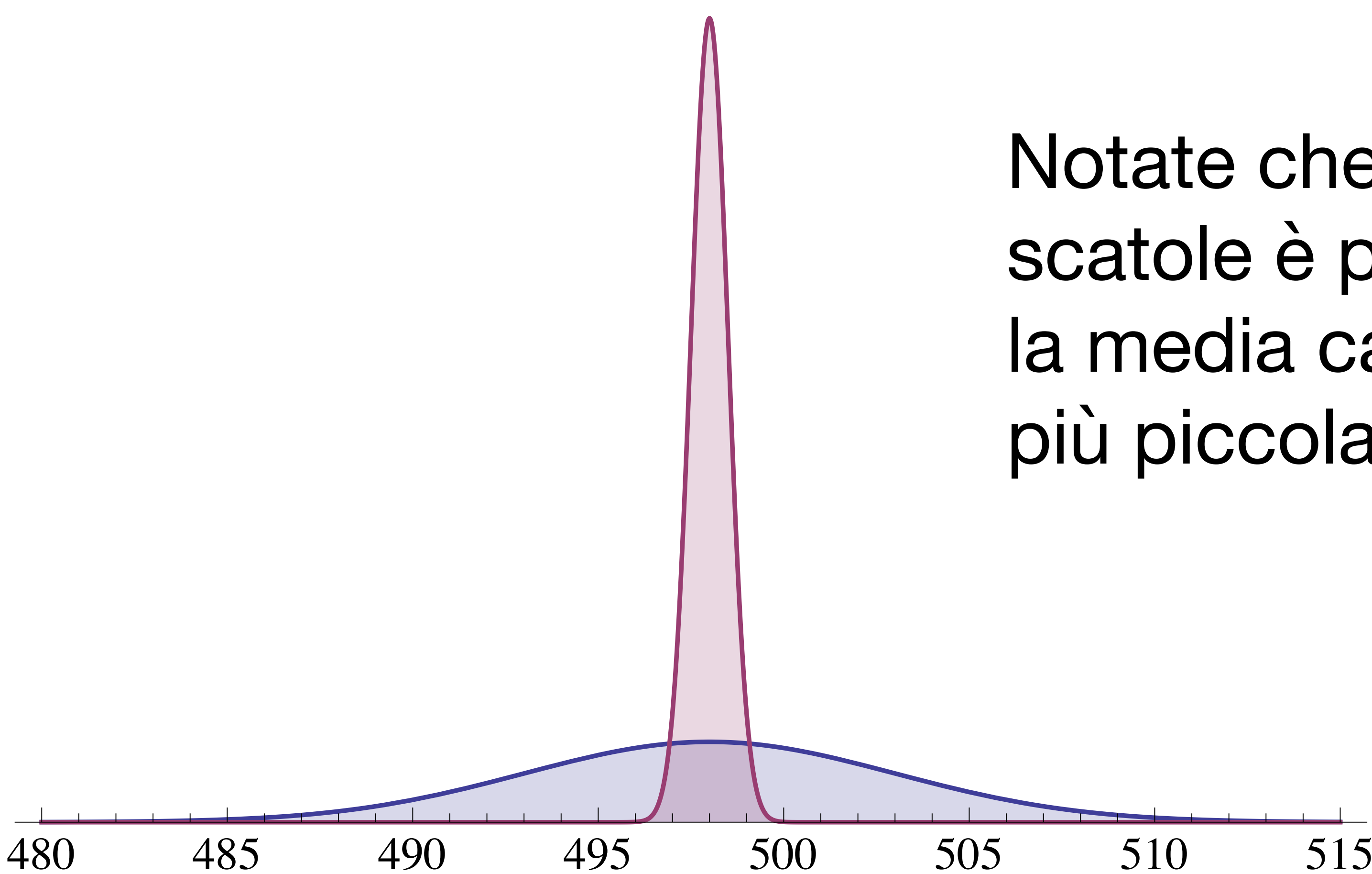
Se il produttore prende un campione di 100 scatole (4 volte il precedente) **qual'è l'errore standard?**

La stima che ottengo
da un campione di 100 casi
ha un **errore standard di 5/10 = 0.5g**



Esempio - aumento della dimensione

L'errore si dimezza con un campione 4 volte più grande

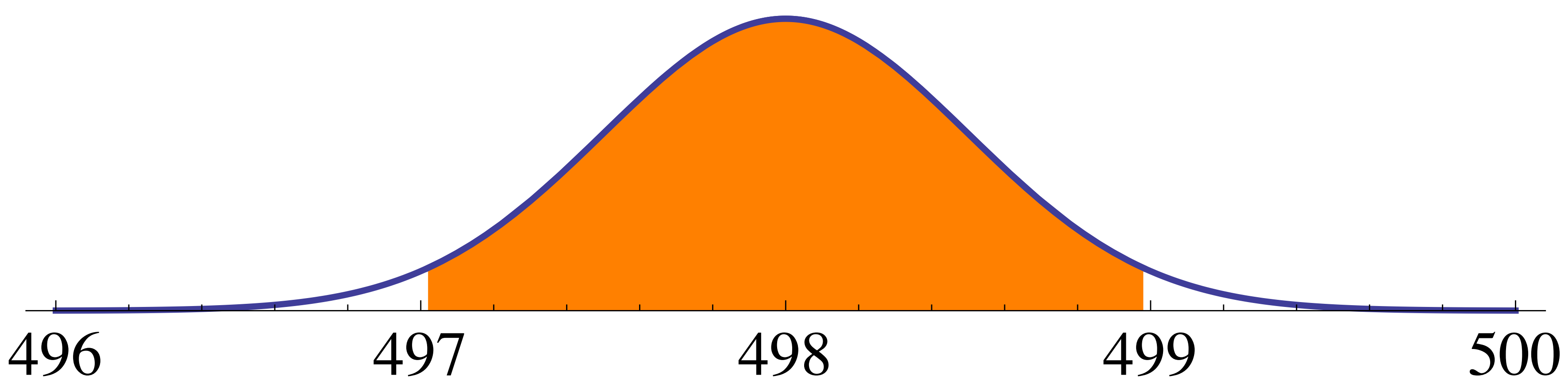


vera media = 498g

Notate che con un campione di 100 scatole è praticamente sicuro che la media campionaria verrà sempre più piccola di 500g

Intervallo di accettazione

È l'intervallo centrato sulla media della popolazione che con probabilità 95% contiene la media campionaria



$$P\left[\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right] = 95\%$$

498 - 1.96 * 0.5 = **497g** 498 + 1.96 * 0.5 = **499g**

Stimatore media campionaria

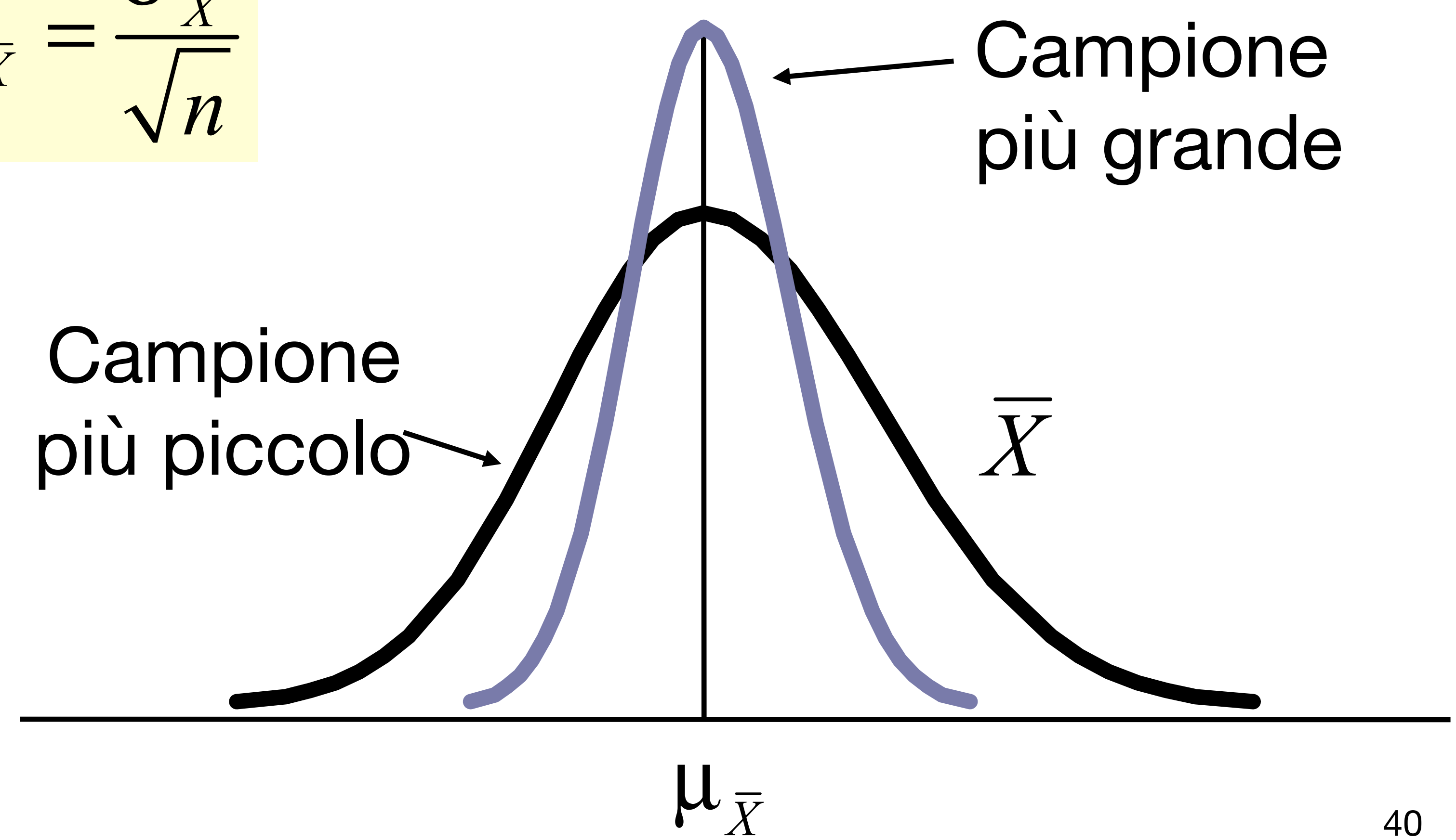
La media campionaria è uno **stimatore** della media della popolazione

Nel campionamento ripetuto lo stimatore

1) ha una media uguale alla media della popolazione $\mu_{\bar{X}} = \mu_X$

2) ha un errore standard $\sigma_{\bar{X}} = \dfrac{\sigma_X}{\sqrt{n}}$

L'errore è minore se il campione e più grande



Errore standard della media

In generale, l’errore standard della media campionaria è

- direttamente proporzionale alla deviazione standard della popolazione. **Quanto più la popolazione è variabile, tanto più la media varia da campione a campione**
- inversamente proporzionale alla radice della dimensione del campione. Quanto più grande è il campione, tanto meno la media varia da campione a campione

I valori grandi e piccoli **si compensano** e la media è meno variabile delle singole osservazioni

Stimatore proporzione

La proporzione campionaria è uno **stimatore** della probabilità di successo nella popolazione

Nel campionamento ripetuto lo stimatore X/n

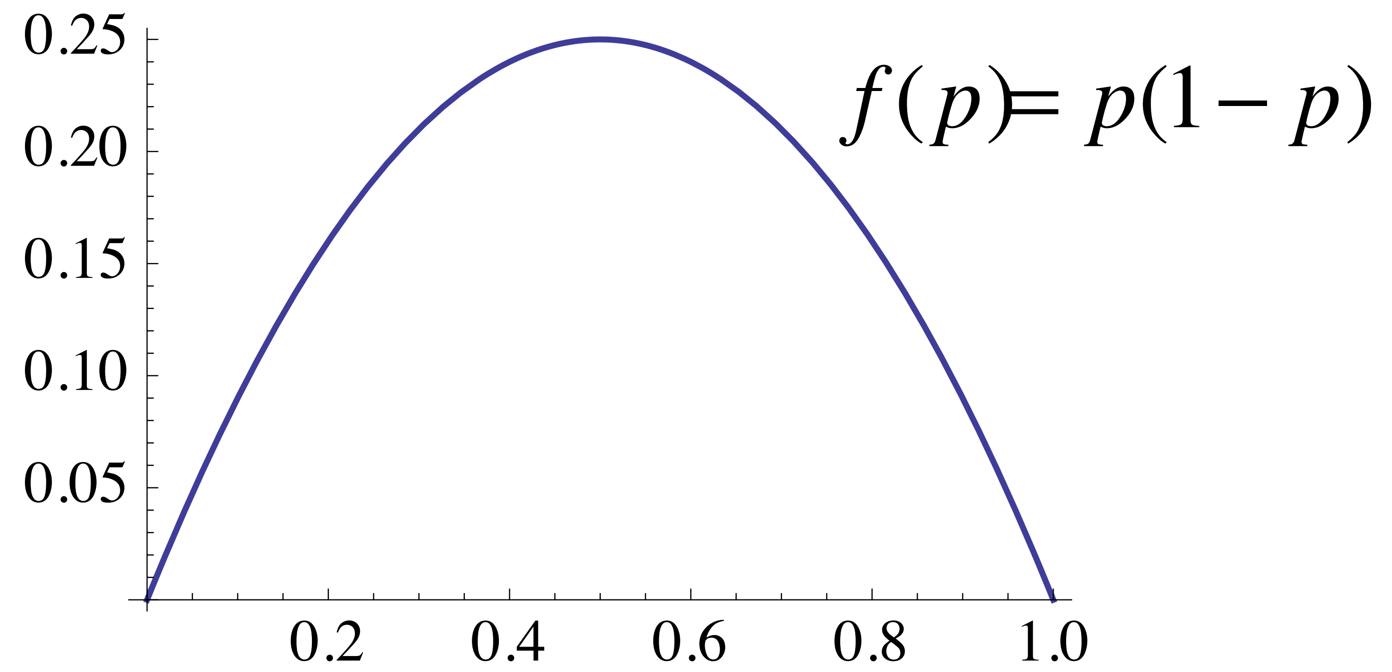
1) ha una media uguale alla media della popolazione $E(X/n) = p$

2) ha un errore standard $\sigma(X/n) = \sqrt{pq/n}$

Errore standard della proporzione

In generale, l'errore standard della proporzione è

- direttamente proporzionale a *pq*. **Quanto più p è vicino a 0.5, tanto più la proporzione varia da campione a campione**
- inversamente proporzionale alla radice della dimensione del campione. Quanto più grande è il campione, tanto meno la proporzione varia da campione a campione



Proporzione campionaria: limite superiore dell'errore standard

La deviazione std della Bernoulli ha un massimo quando p=0.5

$$\sqrt{p(1-p)} = \sqrt{0.5(1-0.5)} = 0.5$$

Quindi al massimo l'errore standard della proporzione campionaria può essere

$$\sigma_{\hat{P}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}} = \frac{1}{\sqrt{4n}}$$

Esempio

In un campione di n elementi per un sondaggio d'opinione (favorevole/contrario)

il massimo errore che si può commettere per stimare p è

$n =$ **25**: 0.10 (cioè **10**%)
$n =$ **100**: 0.05 (cioè **5**%)
$n =$ **2500**: 0.01 (cioè **1**%)

Teorema centrale del limite

La distribuzione campionaria della media è normale anche se la popolazione non è normale **purchè il campione sia grande.**

Sia

$$(X_1, X_2, \ldots, X_n)$$

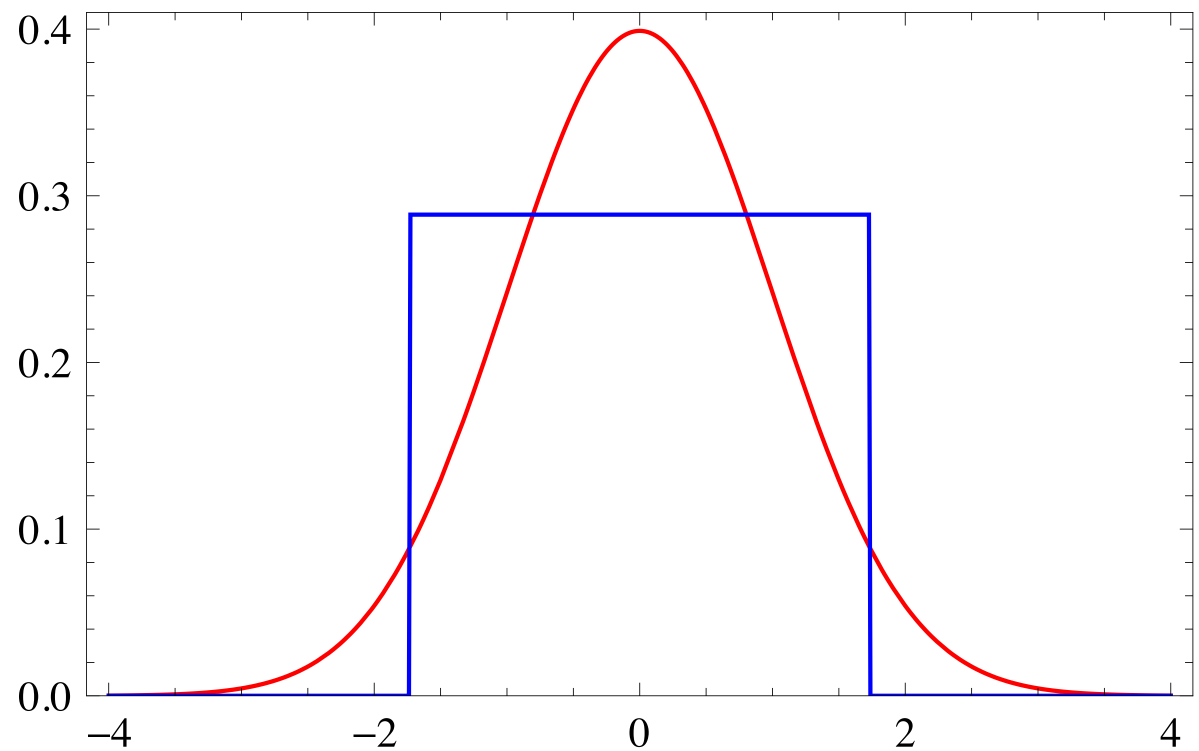
un campione casuale da una popolazione con distribuzione di probabilità *X* **qualsiasi anche incognita**

Sia n > = 30

Allora la media campionaria $\bar{X} \approx N(\mu, \sigma^2/n)$

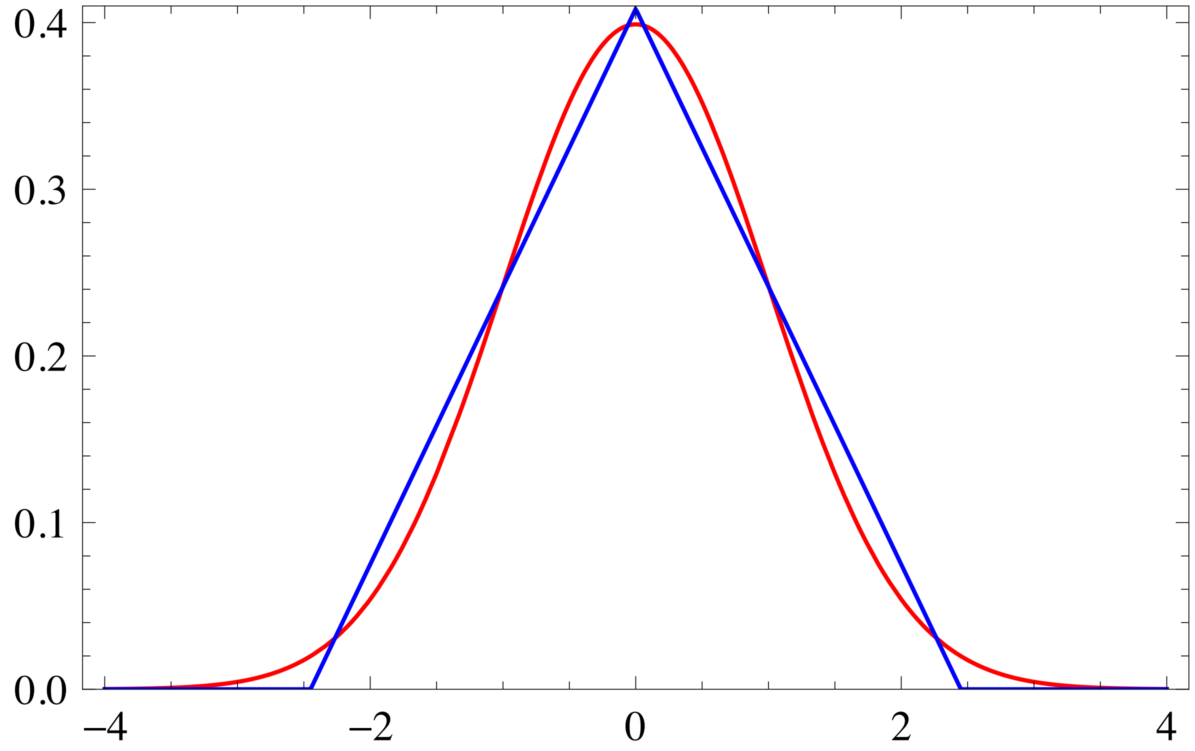
Esempio: Popolazione uniforme

n = 1



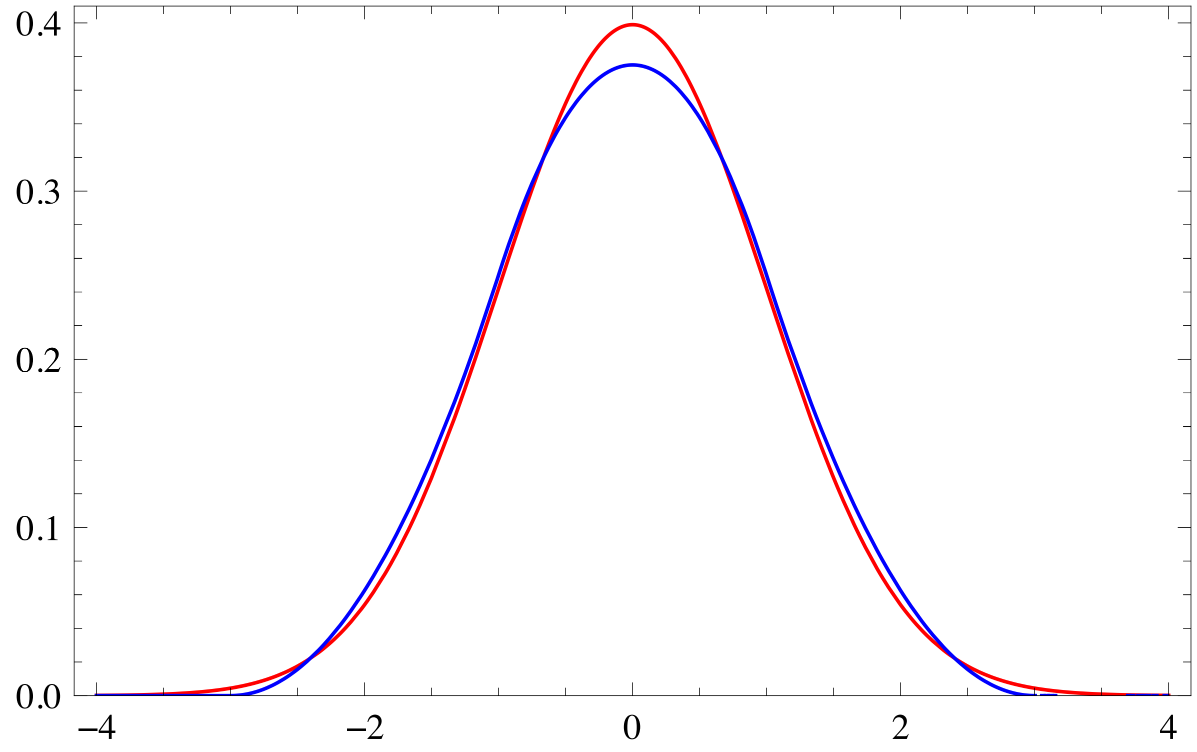
Esempio: Popolazione uniforme

n = 2



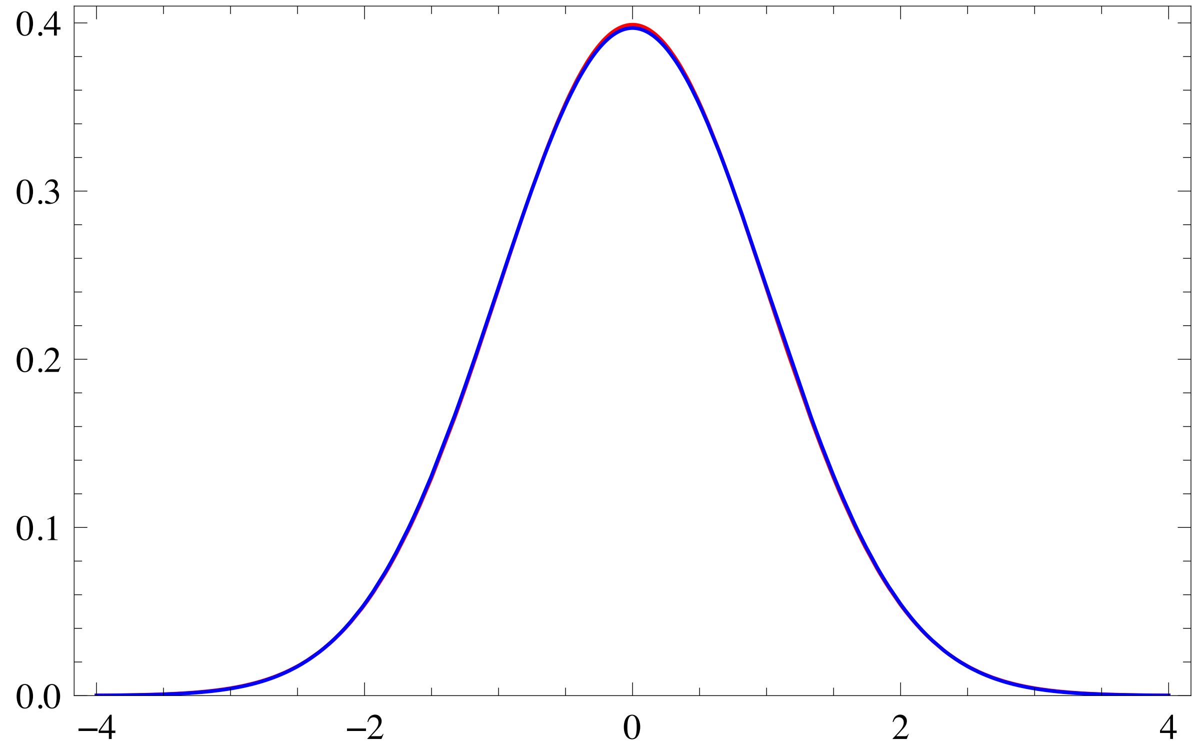
Esempio: Popolazione uniforme

n = 3



Esempio: Popolazione uniforme

n = 30



Applicazione del TLC

Campione da una Distribuzione di Bernoulli

$$(X_1, \ldots, X_n)$$

Il numero di successi nel campione è $X_1 + \cdots + X_n$

La proporzione di successi nel campione è

$$\hat{P} \equiv (X_1 + \cdots + X_n)/n$$

cioè la media campionaria!

1) La sua distribuzione **esatta** è Binomiale (divisa per n)
2) Per il TCL si **approssima** con una normale $\hat{P} \approx N(p, pq/n)$
   se n > 30 e npq > 9

Esempio

Supponiamo che il 75% di tutti i potenziali clienti
di un centro commerciale sia soddisfatto del servizio.
La popolazione si descrive come una Bernoulli con p=0.75

Si estrae un campione di n = 200 clienti.

L'esatto errore standard della proporzione $\hat{P}$ è

$$\sigma_{\hat{P}} = \sqrt{\frac{0.75(1-0.75)}{200}} = 0.0306$$

Qual è la probabilità di osservare un campione in cui i clienti soddisfatti sono < 70%?

Calcoli

Qui si usa il CLT

perchè $n = 200$ e $nqp = 37.5 > 9$

$$\Pr[\hat{P} < 0.7] \approx \Pr[X < 0.7; X \sim N(0.75, 0.75 \cdot 0.25/200)]$$

$$= \Pr\left[Z < \frac{0.7 - 0.75}{0.0306}\right]$$

$$= \Pr[Z < -1.633] = 0.051$$