

Lezione 7

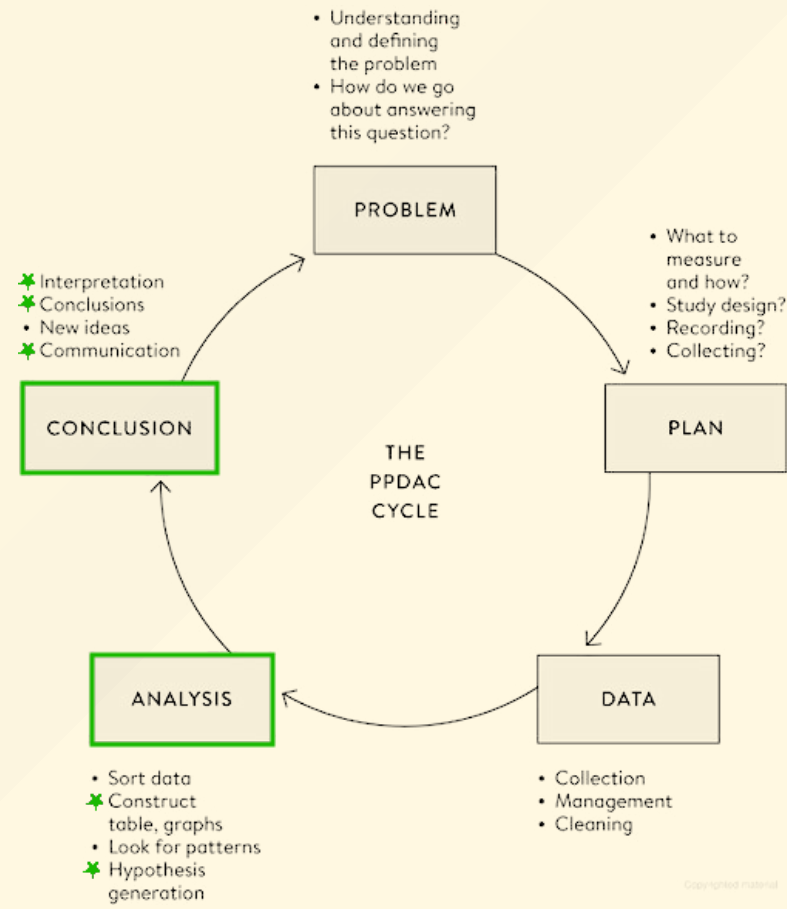
La statistica inferenziale

(Parte I: Stime e intervalli di confidenza)

Obiettivi di apprendimento

- Saper passare da una distribuzione empirica alla quella di popolazione
- Saper comunicare l'incertezza di una statistica
- Saper calcolare e interpretare un intervallo di confidenza

Le fasi della ricerca

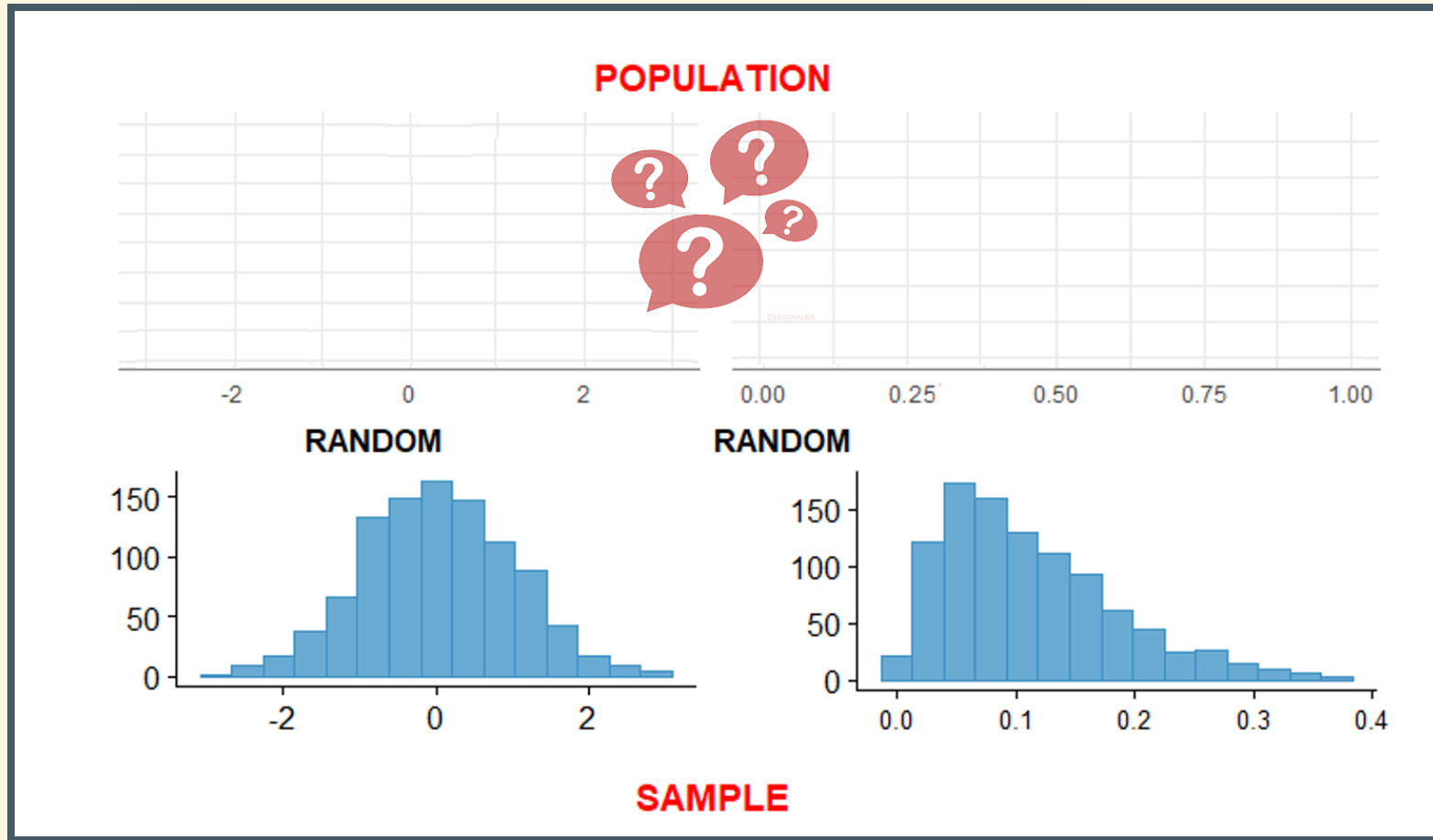


! **Attenzione** !

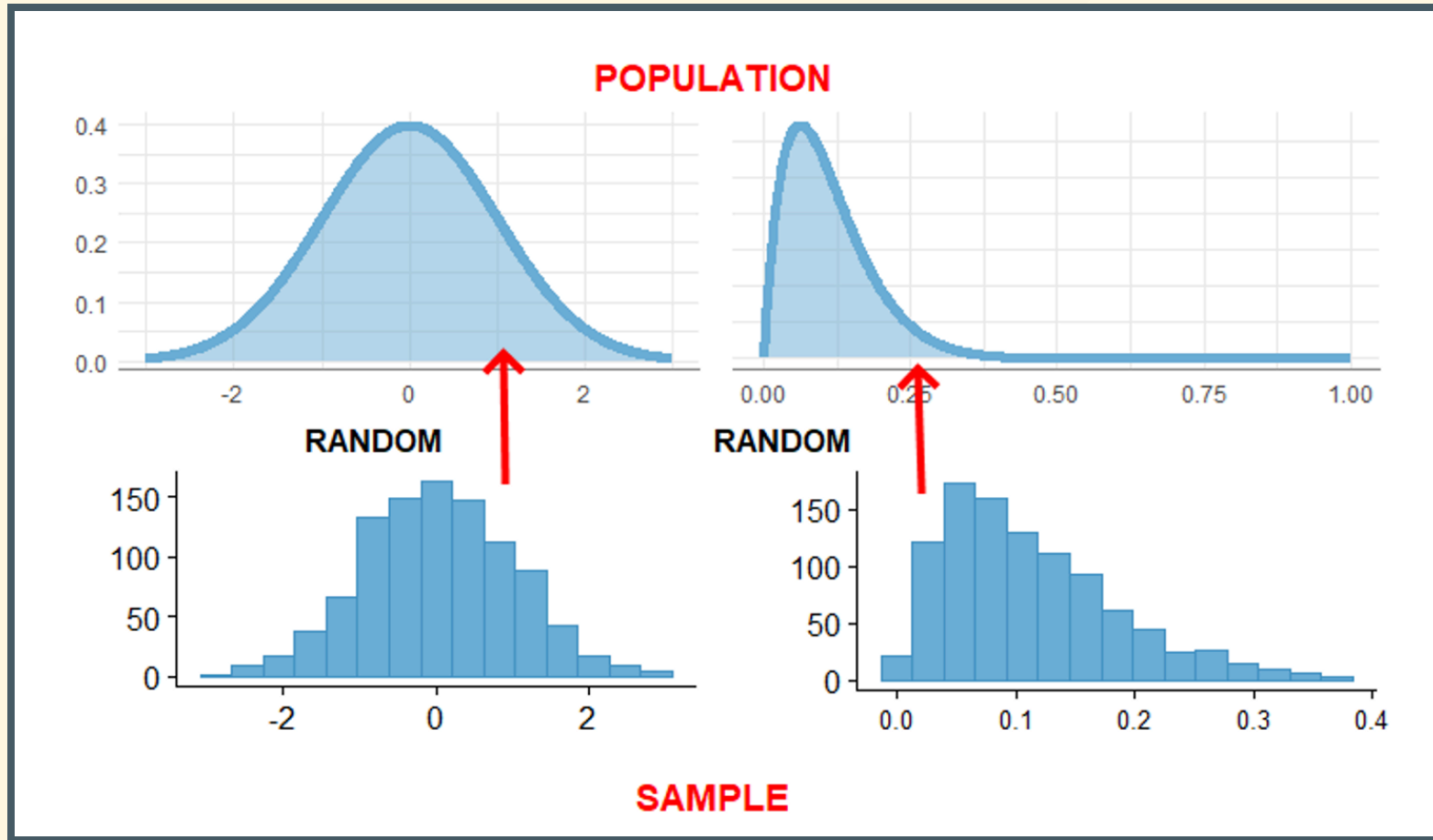
Se questa parte vi sembra difficile è perché è difficile.

Potreste doverci spendere un bel po' di tempo prima di riuscire a capirla del tutto: non vi preoccupate, è normale e ci siamo passati tutti!

Dalla campione alla popolazione



Dalla campione alla popolazione



Quanto siamo precisi?

“ Quanti partner sessuali le persone in Gran Bretagna riferiscono di aver avuto nella loro vita? ”

$$N_{\text{donne}} = 1100$$
$$N_{\text{uomini}} = 796$$

	Uomini 35-44	Donne 35-44
Moda	1	1
Range	0-500	0-550
Media	14.3	8.5
SD	24.2	19.7
Mediana	8	5
IQR	4-18 (14)	3-10 (7)

Ricalcoliamo senza i "valori estremi"

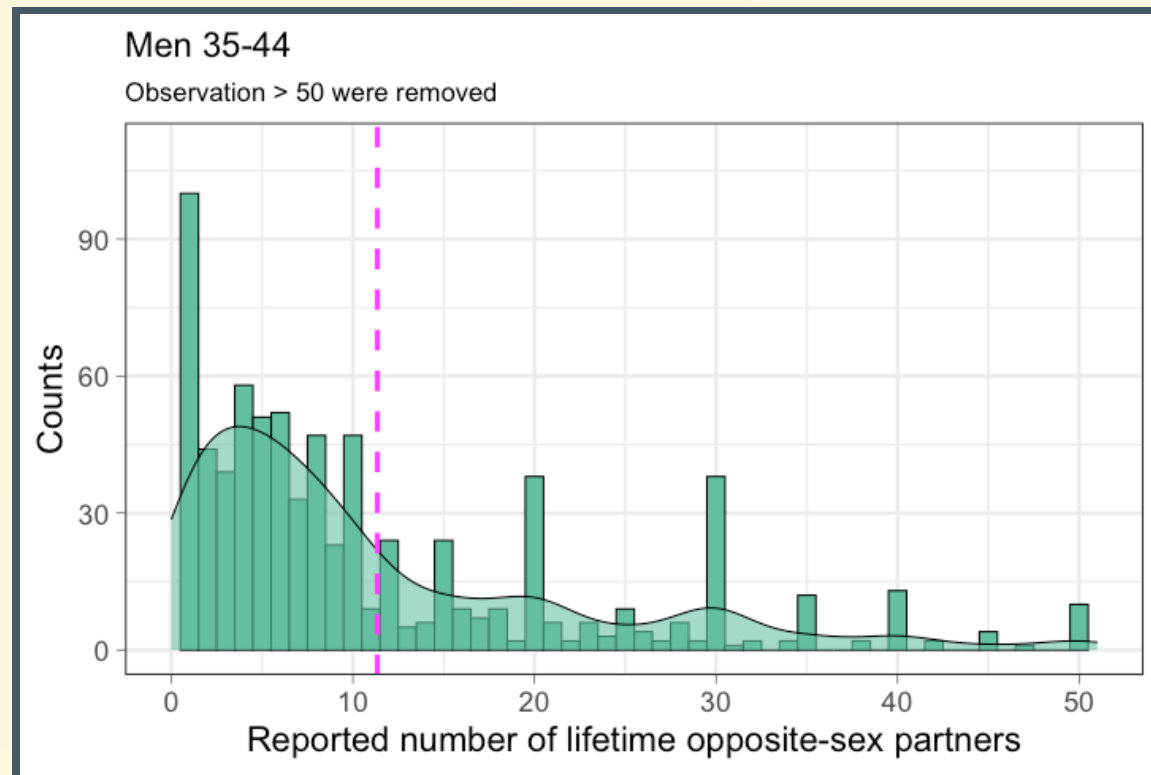
“ Quanti partner sessuali gli uomini inglesi, tra i 35 e 44 anni di età, riferiscono di aver avuto nella loro vita? ”

$$N_{\text{uomini}} = 760$$

Uomini 35-44	
Moda	1
Range	0-50
Media	11.4
SD	11.2
Mediana	7
IQR	4-16 (12)

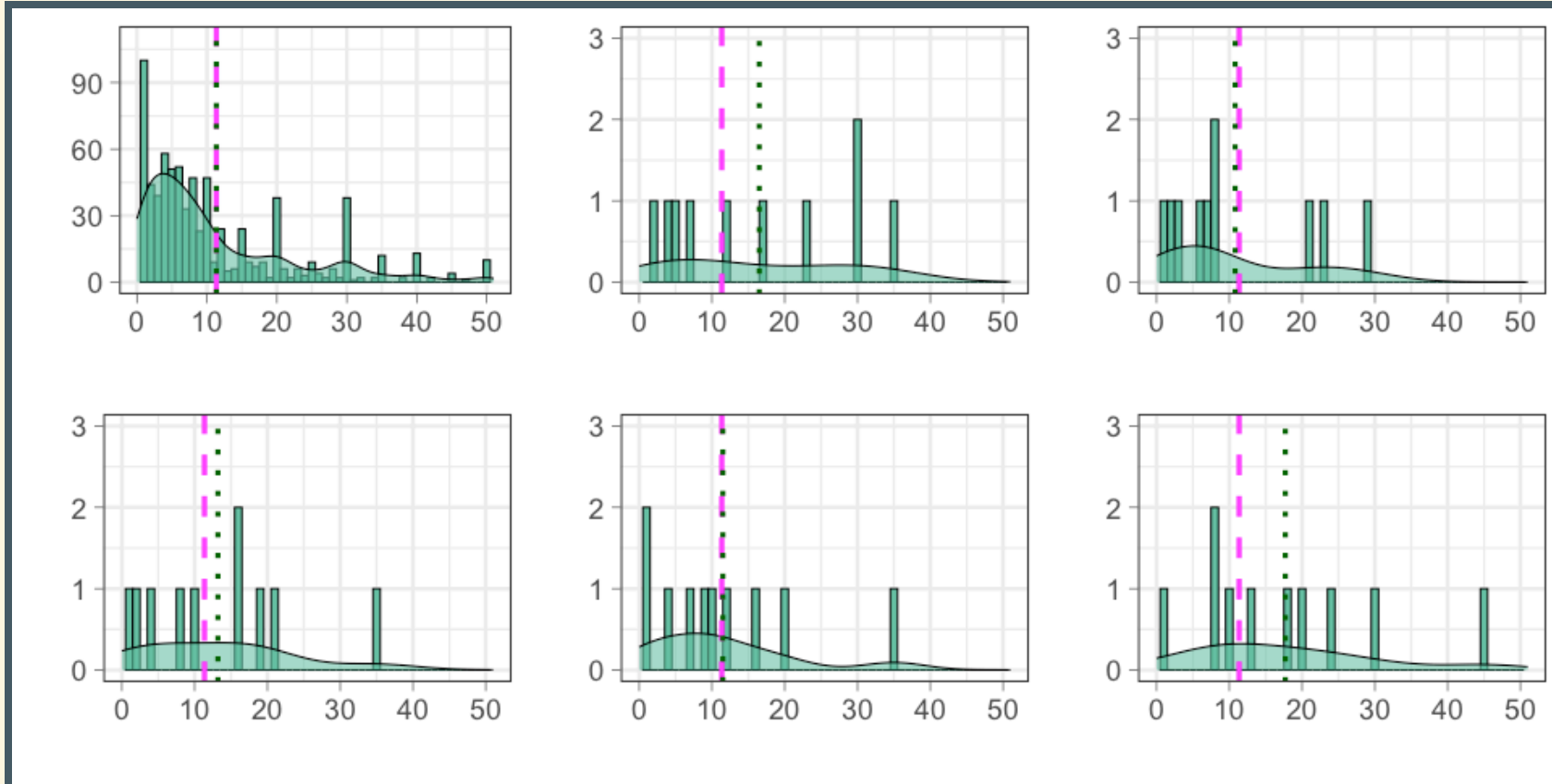
La dimensione del campione

“ Quanti partner sessuali gli uomini inglesi, tra i 35 e 44 anni di età, riferiscono di aver avuto nella loro vita? ”



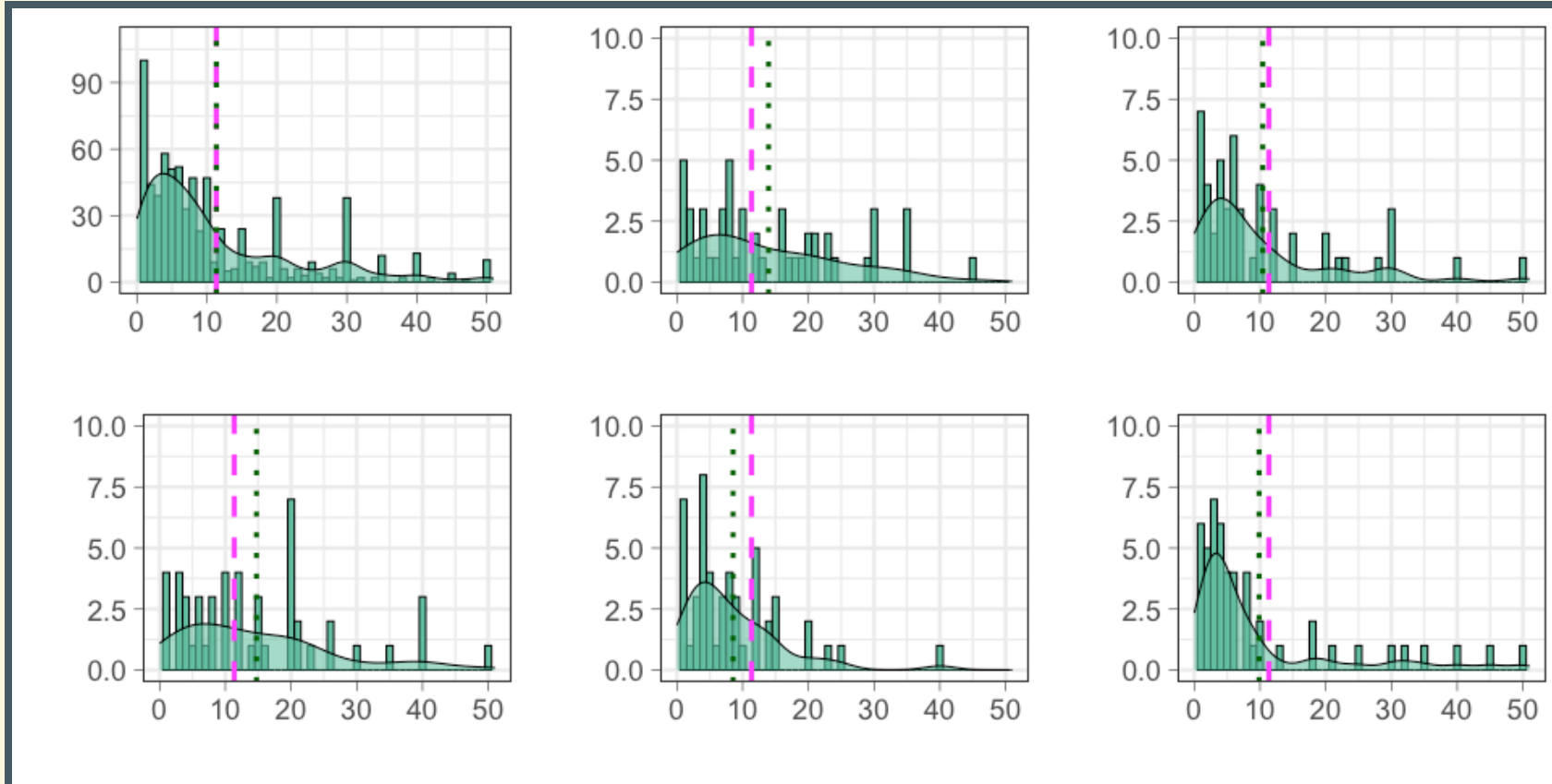
La dimensione del campione

$$N_{\text{campione}} = 10$$



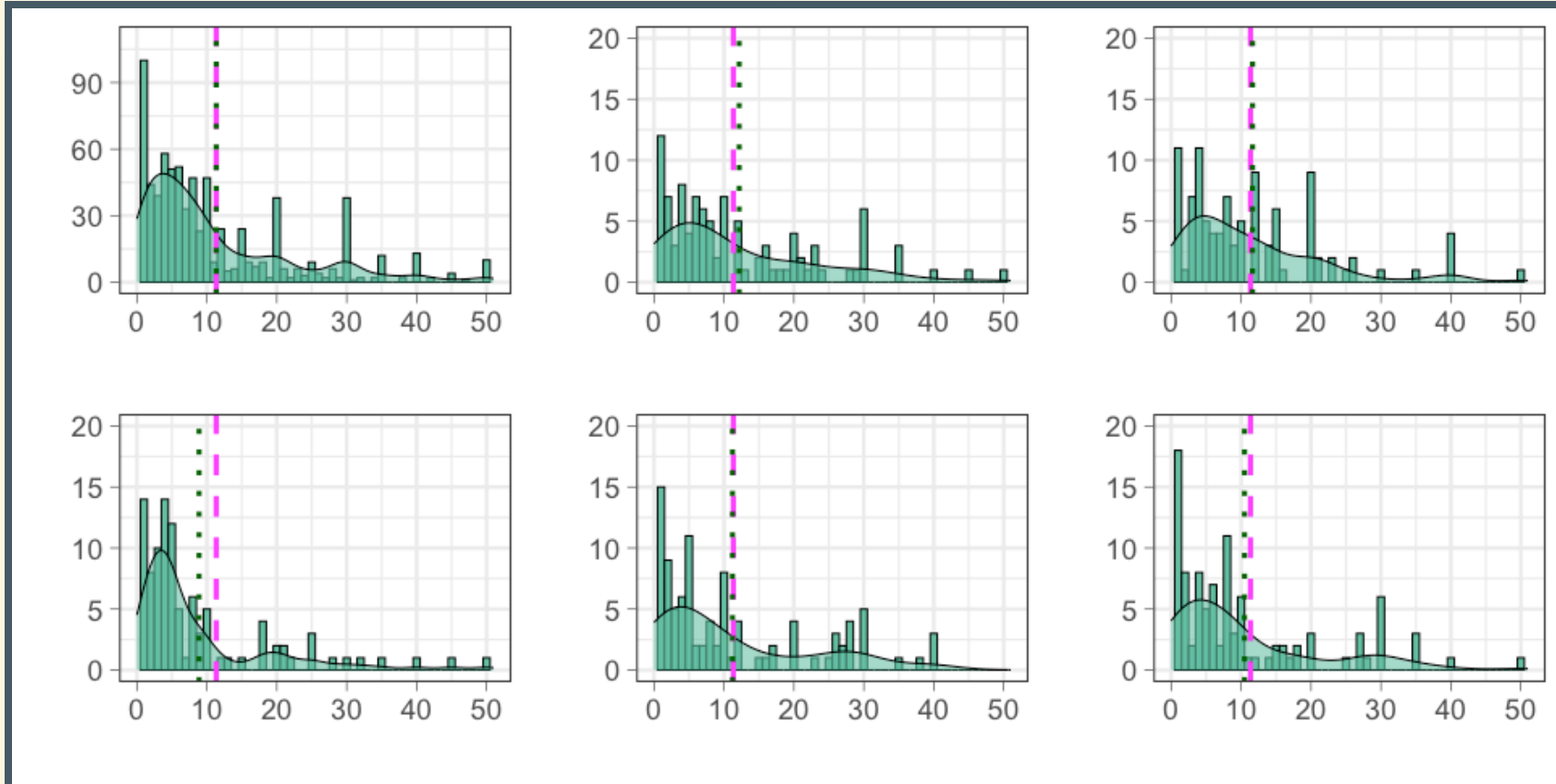
La dimensione del campione

$$N_{\text{campione}} = 50$$



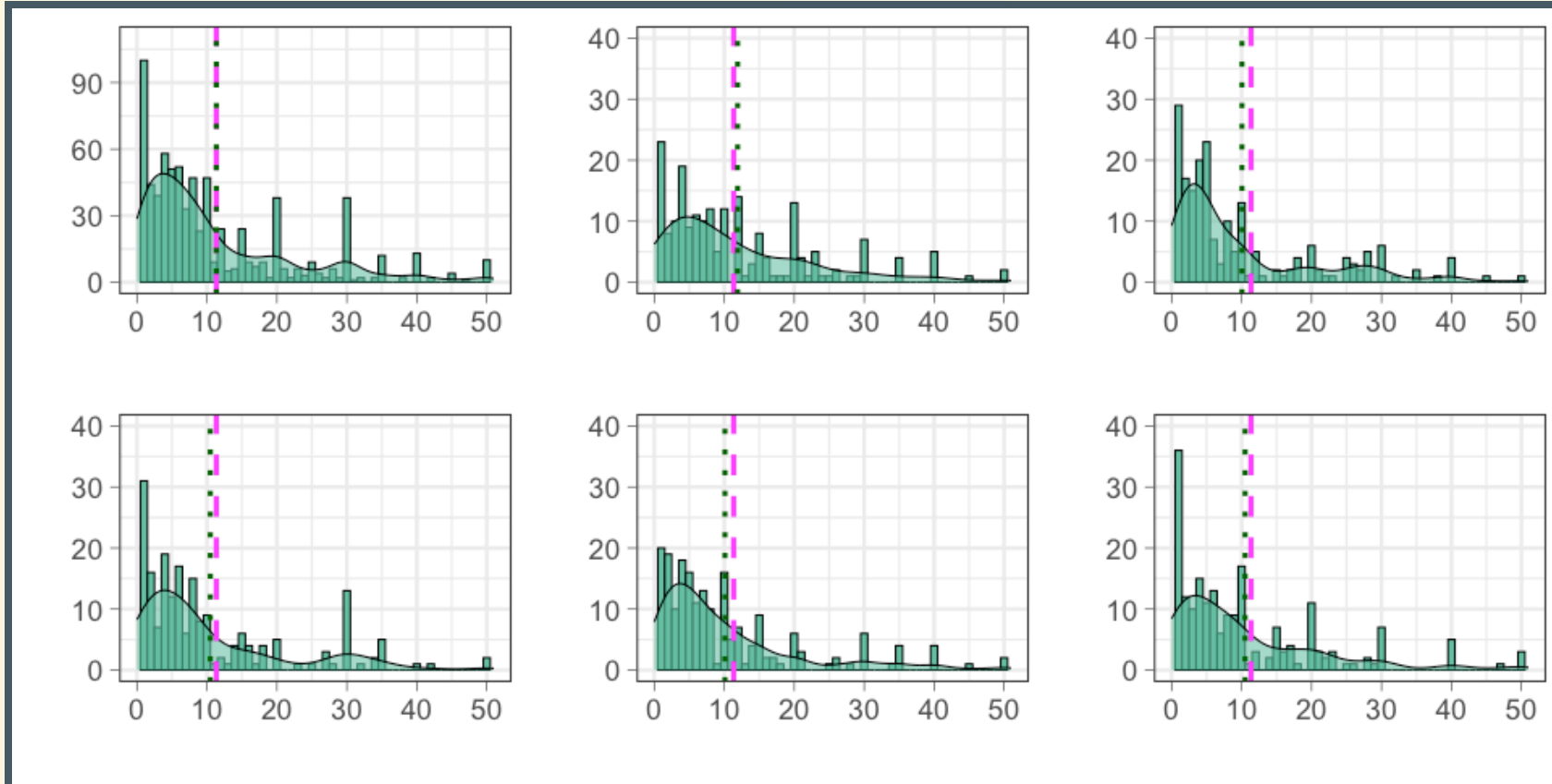
La dimensione del campione

$$N_{\text{campione}} = 100$$



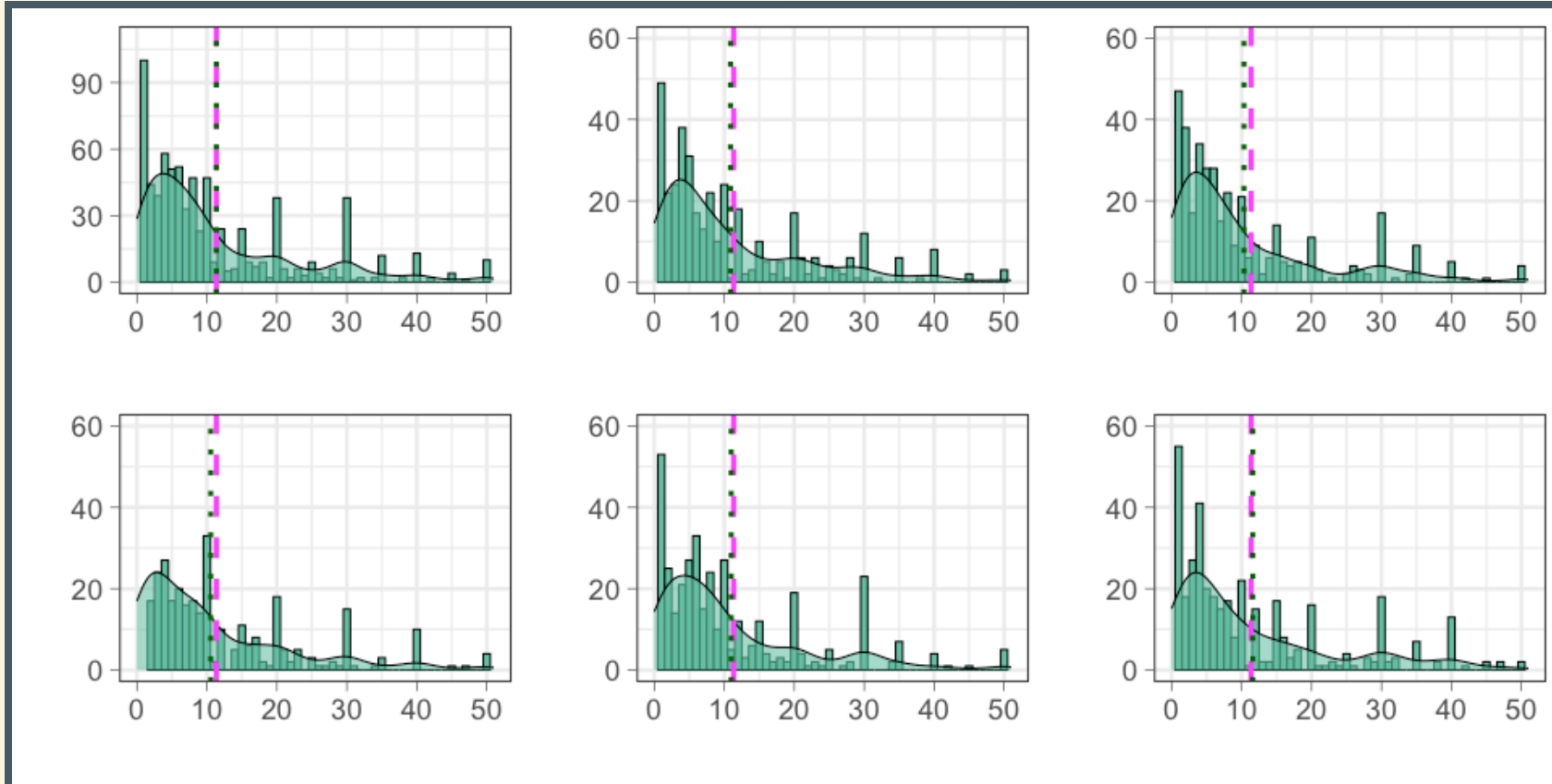
La dimensione del campione

$$N_{\text{campione}} = 200$$



La dimensione del campione

$$N_{\text{campione}} = 380$$



Esercizio #1

- ?
- Al crescere della dimensione del campione
- a) migliorano le stime dei parametri
 - b) le stime dei parametri diventano più sensibili alle singole osservazioni
 - c) non c'è differenza
 - c) non ho abbastanza elementi per rispondere

Esercizio #1 -- Soluzione

- ?
- Al crescere della dimensione del campione
- a) migliorano le stime dei parametri ☒
 - b) le stime dei parametri diventano più sensibili alle singole osservazioni
 - c) non c'è differenza
 - c) non ho abbastanza elementi per rispondere

Quanto siamo precisi?

Con questo esempio, abbiamo introdotto due concetti:

1. Campioni più grandi stimano meglio i parametri di una popolazione
2. Continuare ad estrarre campioni ci dà un'idea della variazione attorno al valore "plausibile" del parametro che ci interessa

Quindi come procediamo?

Stima dei parametri e del margine di errore

Come stimo la variazione rispetto al valore reale nella popolazione se quello che sto cercando è proprio il valore reale nella popolazione?

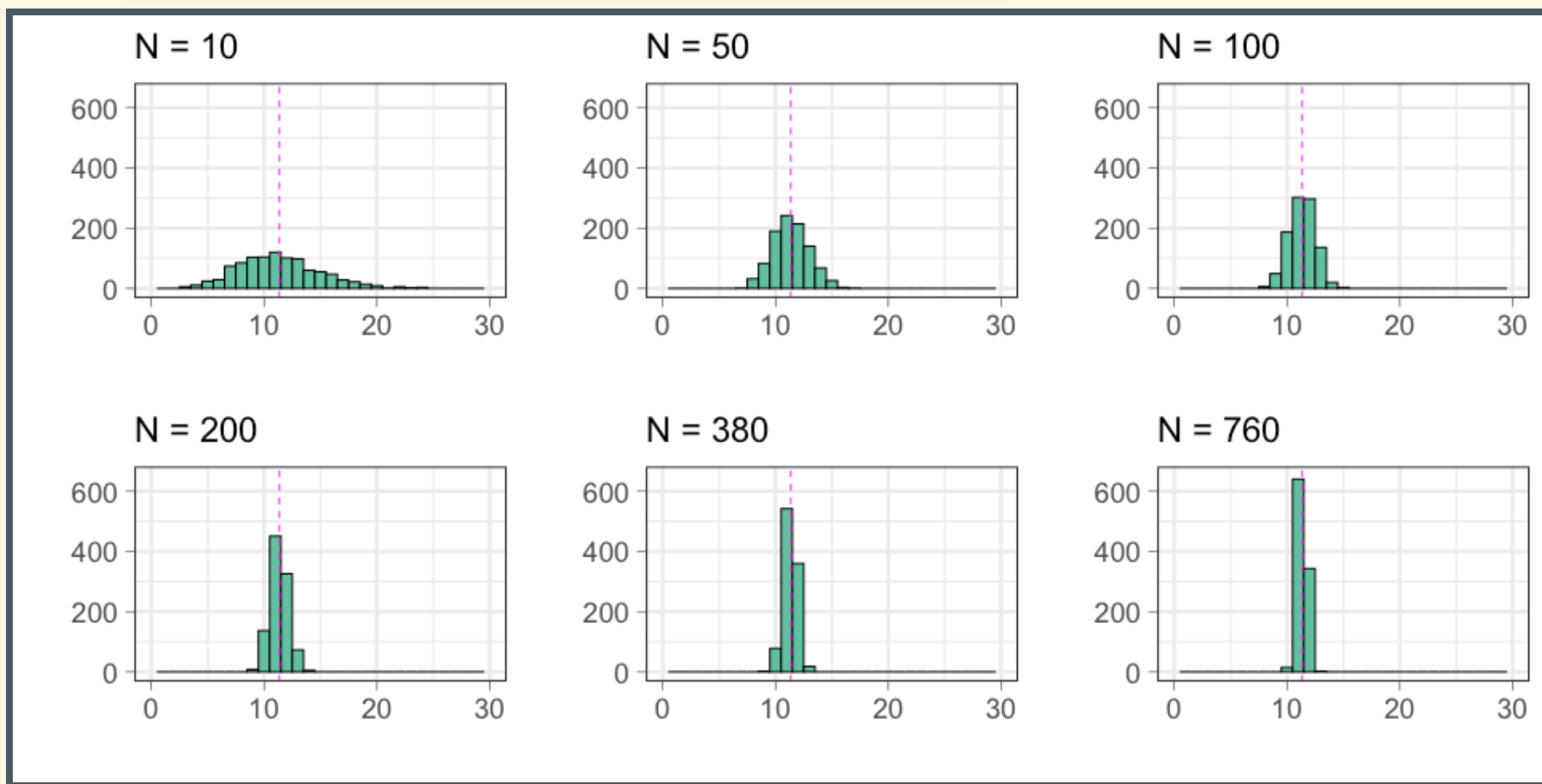


Stima dei parametri e del margine di errore

1. Assumendo che la popolazione assomigli al campione
→ via bootstrapping
2. Facendo assunzioni matematiche sulla forma della distribuzione nella popolazione
→ via distribuzione campionaria & teorema del limite centrale

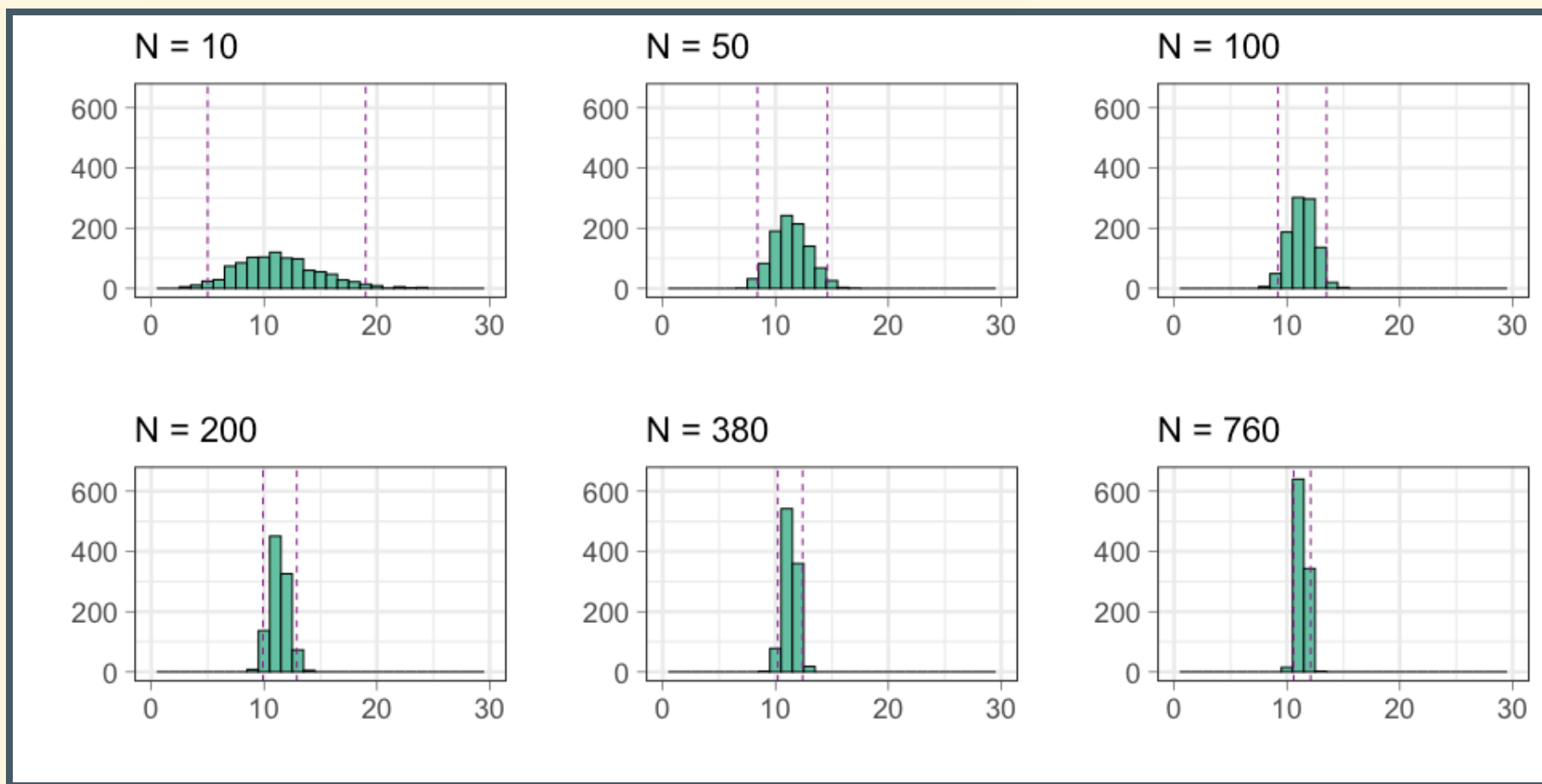
Stima dei parametri e del margine di errore

$$N_{\text{Bootstrapping}} = 1000$$



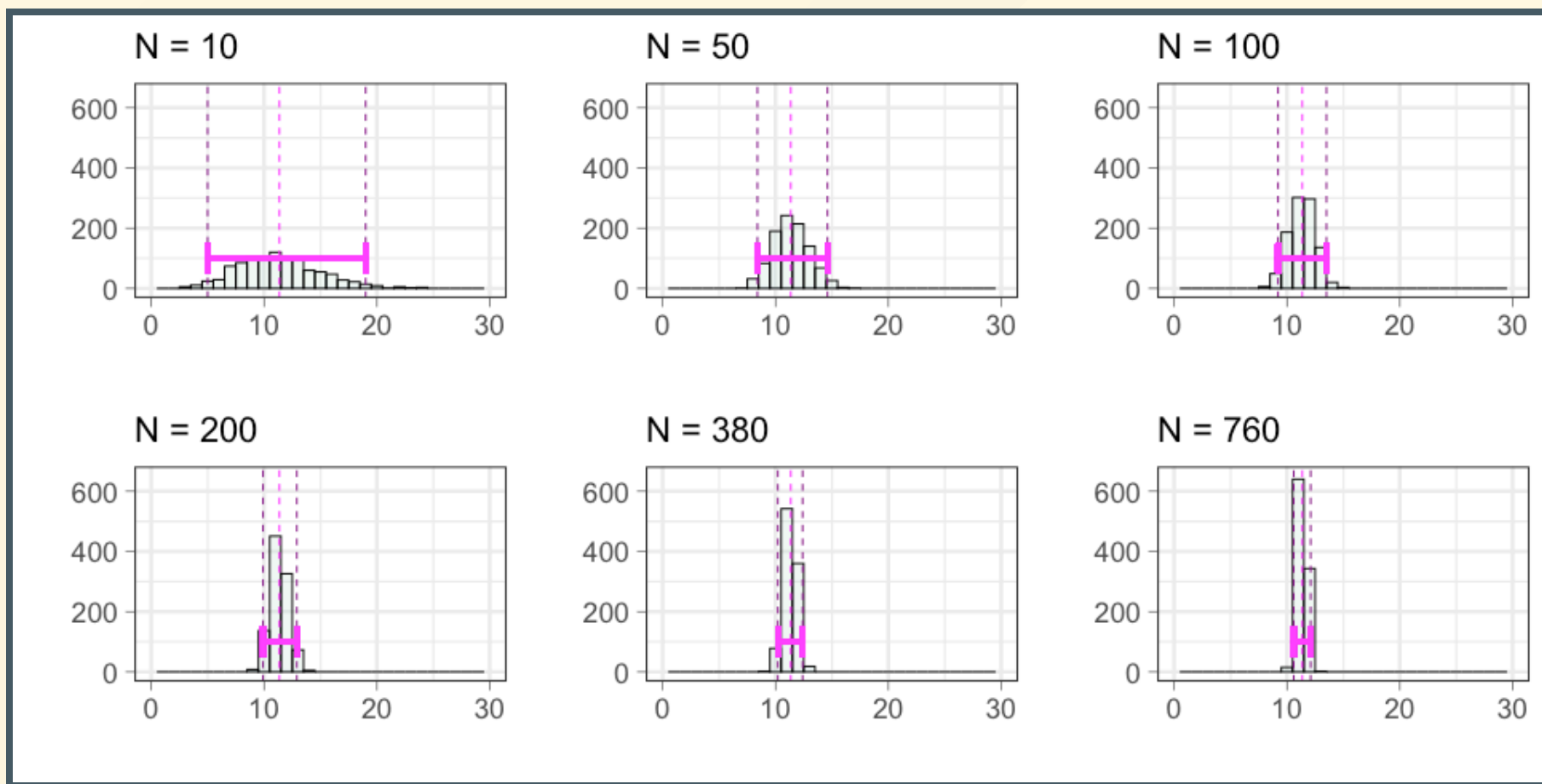
Il margine di errore (o intervallo di confidenza)

Intervallo che contiene il 95% delle medie ottenute via Bootstrapping



Il margine di errore (o intervallo di confidenza)

Intervallo che contiene il 95% delle medie ottenute via Bootstrapping



Intervallo di confidenza

Intervallo che contiene il 95% delle medie ottenute via Bootstrapping

$$N_{\text{Bootstrapping}} = 1000$$
$$\bar{x} = 11.4$$

N_{campione}	Media	95% CI
10	11.4	(5.0; 19.0)
50	11.4	(8.4; 14.6)
100	11.4	(9.2; 13.5)
200	11.3	(9.9; 12.9)
380	11.3	(10.2; 12.4)
760	11.3	(10.6, 12.1)

Esercizio #2

- ?
- Al crescere del numero di campioni estratti da una popolazione
- a) migliora la stima del parametro
 - b) migliora la stima dell'incertezza del parametro
 - c) non c'è differenza
 - c) non ho abbastanza elementi per rispondere

Esercizio #2 -- Soluzione

? Al crescere del numero di campioni estratti da una popolazione

a) migliora la stima del parametro

b) migliora la stima dell'incertezza del parametro 

c) non c'è differenza

c) non ho abbastanza elementi per rispondere

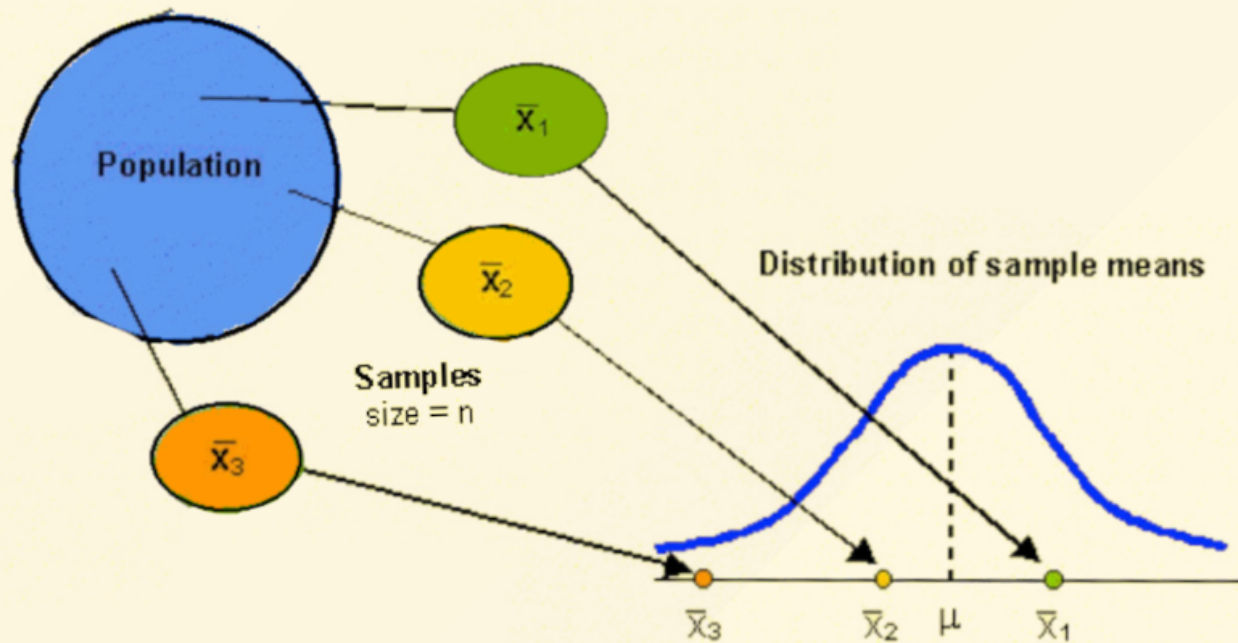
Fermiamoci un attimo

Abbiamo introdotto due concetti difficili e importanti:

1. esiste una variabilità nella stima dei parametri che dipende dal campione
2. la forma della distribuzione delle statistiche non dipende dalla forma della distribuzione originaria e tende alla normale per insiemi grandi

Ora abbiamo le basi per affrontare il secondo approccio per stimare i parametri e l'intervallo di confidenza

La distribuzione campionaria & il teorema del limite centrale



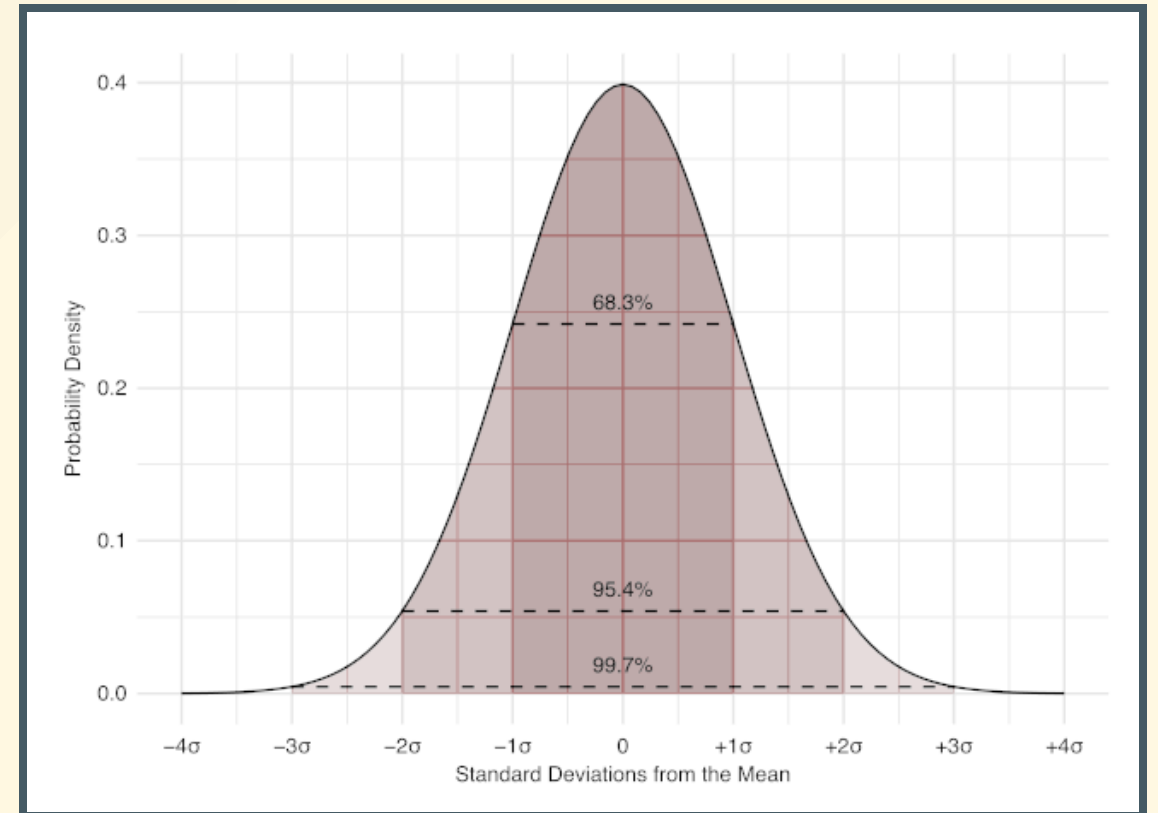
$$\mathcal{N} = \left(\mu, \frac{\sigma^2}{n} \right) \text{ con}$$

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

→ standard error (SE)

Mettiamo i pezzi insieme

- la distribuzione campionaria è una normale
- in una normale, 95% delle osservazioni sono a circa $2 \times \text{SD}$ dalla media
- il nostro intervallo di confidenza (95%) è $2 \times \text{SE}$ dalla media della distribuzione campionaria

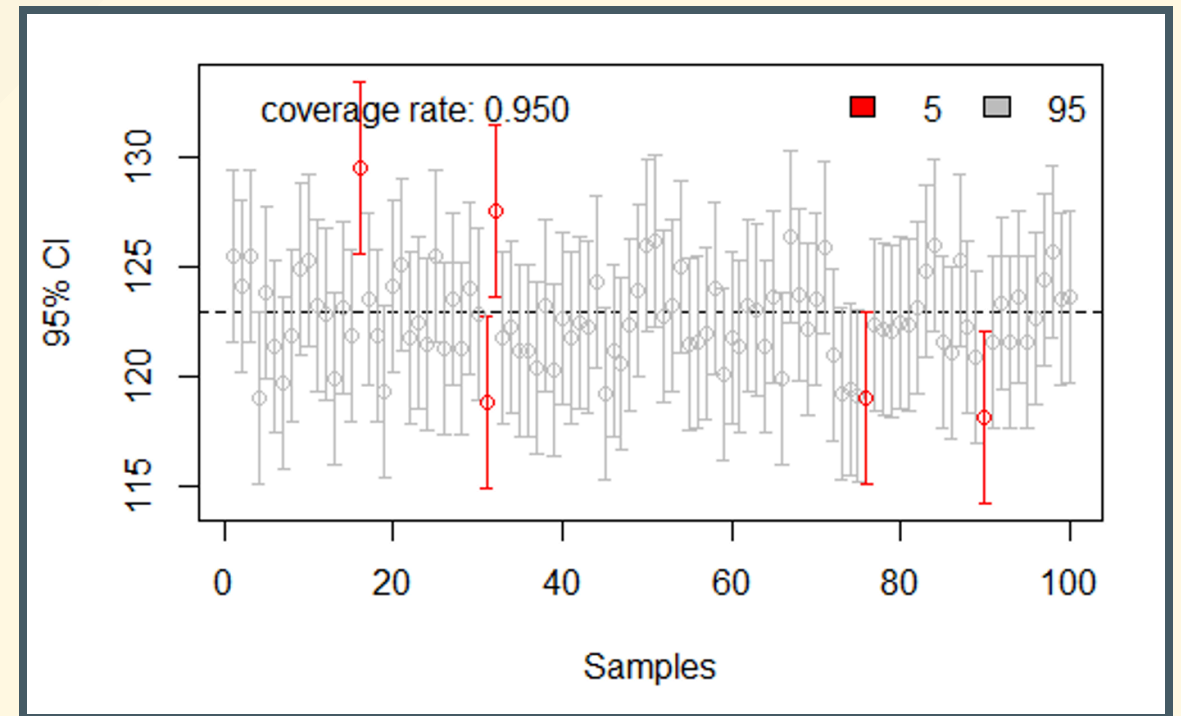


Ma come si interpreta?

Se facessimo 100 campionamenti, 95 stimerebbero un intervallo di confidenza che contiene il vero valore del parametro



Popolazione: Donne italiane dai
25 ai 74 anni
 $\mu = 123$ mmHg




Esercizio #3

- ? Da un sondaggio, risulta che lo stipendio mensile medio di un neolaureato è 1.400€, con un 95% CI = (1.200€ ; 1.600€).
Come interpreto questo risultato?
- a) gli stipendi dei neolaureati sono compresi tra i 1.200 ai 1.600€
 - b) il 95% dei neolaureati riceve tra 1.200 ai 1.600€
 - c) la media degli stipendi dei neolaureati è ragionevolmente compresa tra 1.200 ai 1.600€
 - d) nessuna delle precedenti

Esercizio #3 -- Soluzione

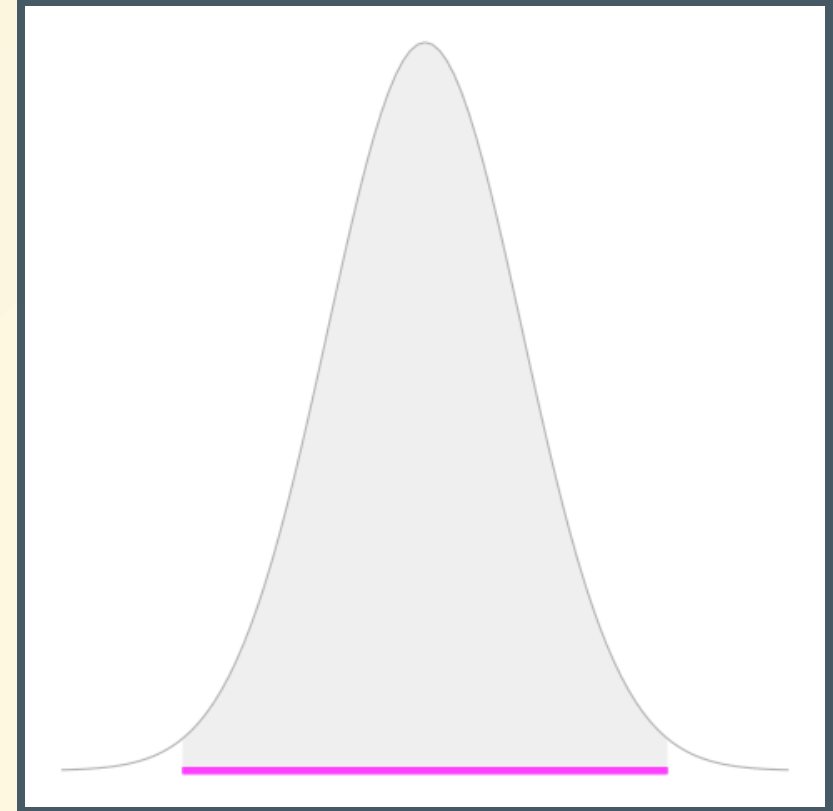
? Da un sondaggio, risulta che lo stipendio mensile medio di un neolaureato è 1.400€, con un 95% CI = (1.200€ ; 1.600€).
Come interpreto questo risultato?

- a) gli stipendi dei neolaureati sono compresi tra i 1.200 ai 1.600€
- b) il 95% dei neolaureati riceve tra 1.200 ai 1.600€
- c) la media degli stipendi dei neolaureati è ragionevolmente compresa tra 1.200 ai 1.600€ 
- d) nessuna delle precedenti

Esercizio #4

? Se l'intervallo di confidenza è largo

- a) è più probabile che includa μ
- b) è meno probabile che includa μ
- c) non c'è differenza
- c) non posso rispondere



Esercizio #4 -- Soluzione

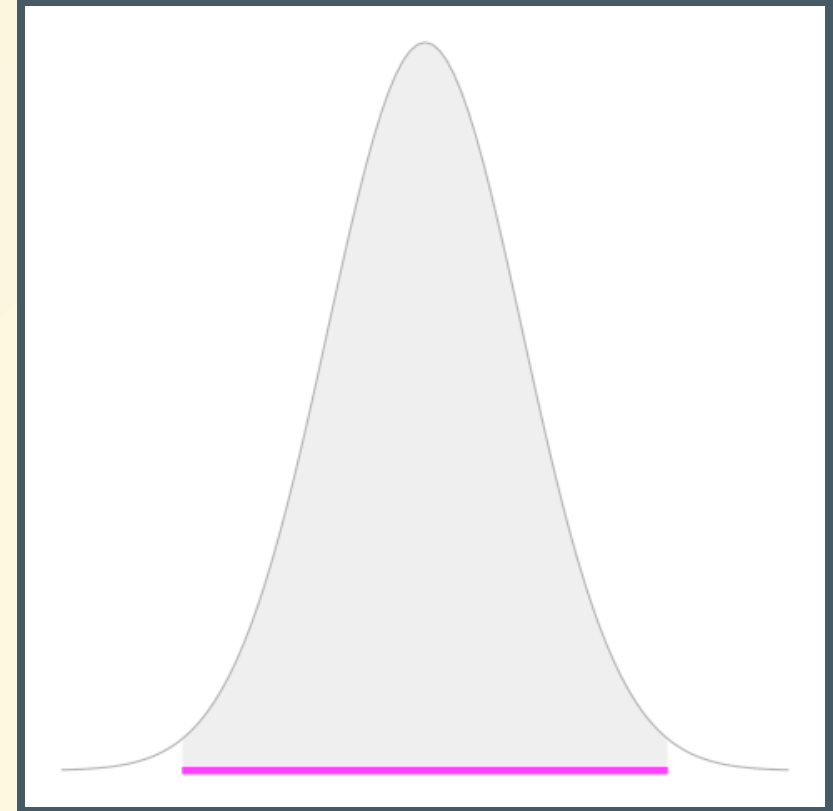
? Se l'intervallo di confidenza è largo

a) è più probabile che includa μ ✓

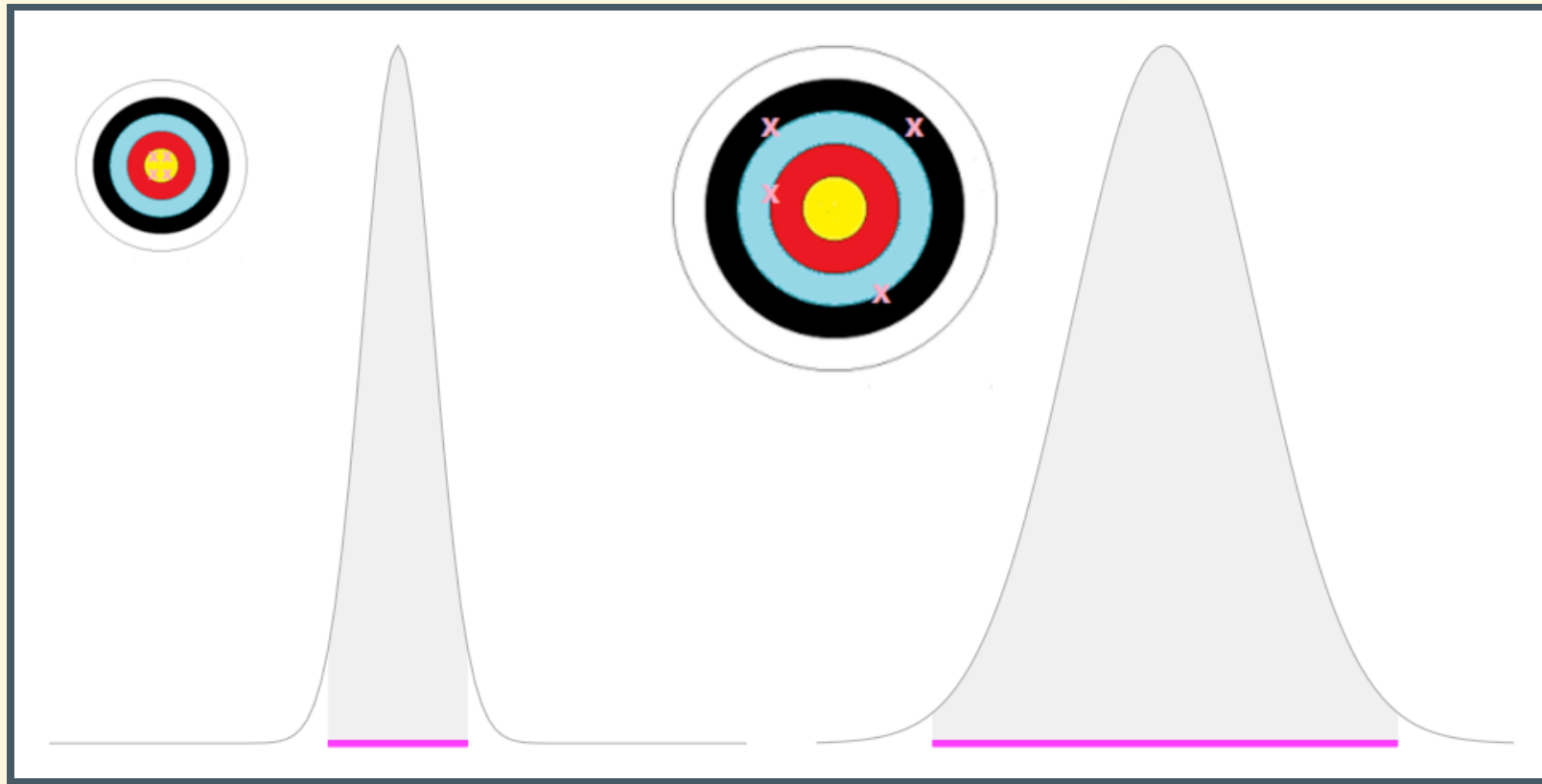
b) è meno probabile che includa μ

c) non c'è differenza

c) non posso rispondere



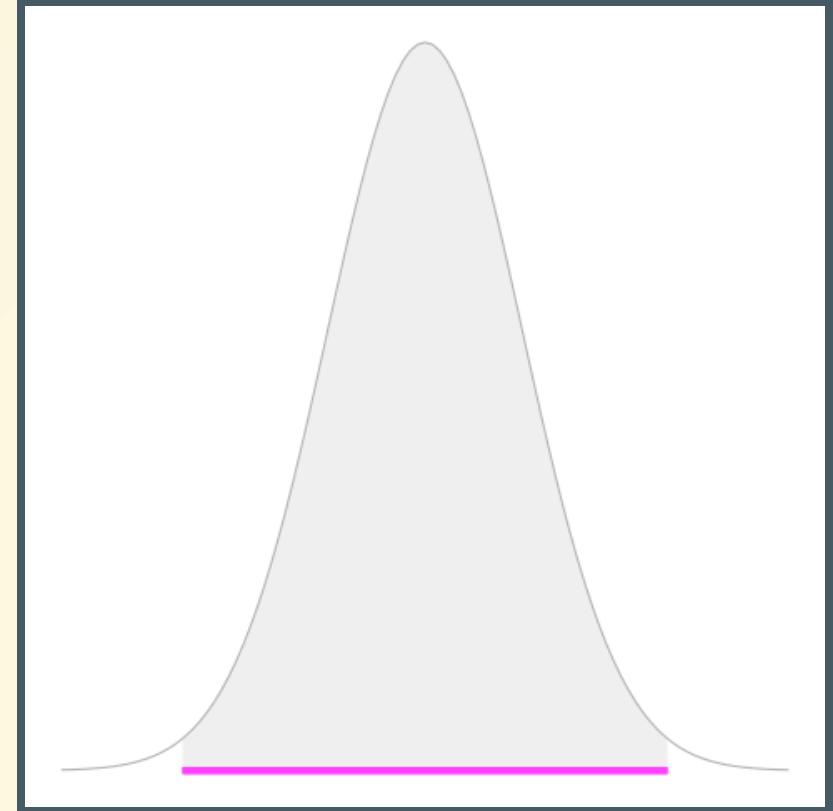
Esercizio #4 -- Soluzione



Esercizio #5

? Più l'intervallo di confidenza è largo

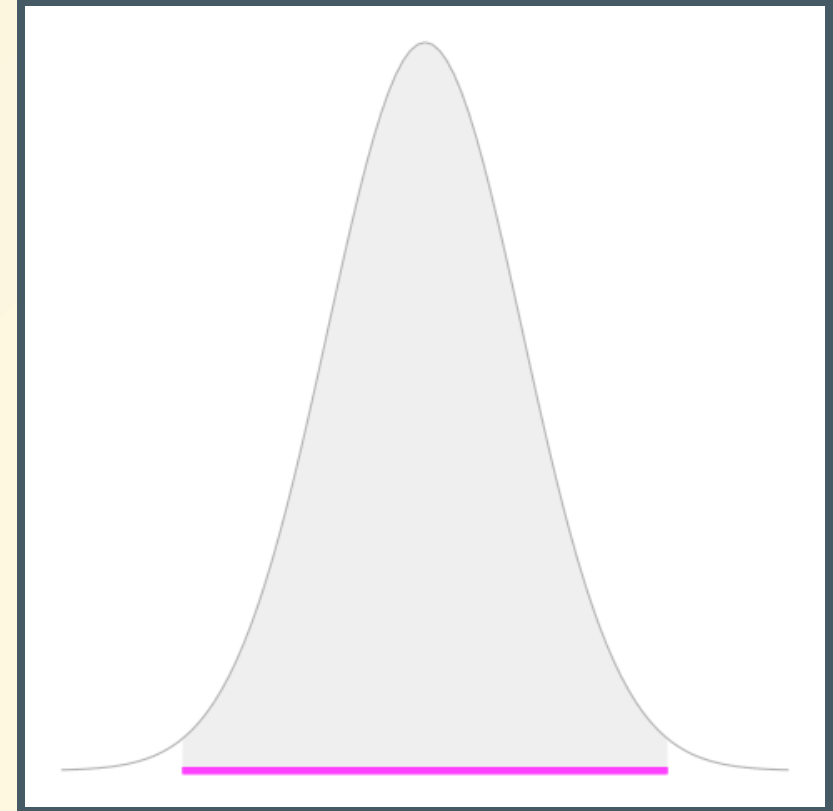
- a) meno siamo precisi
- b) più siamo precisi
- c) non c'è differenza
- c) non posso rispondere



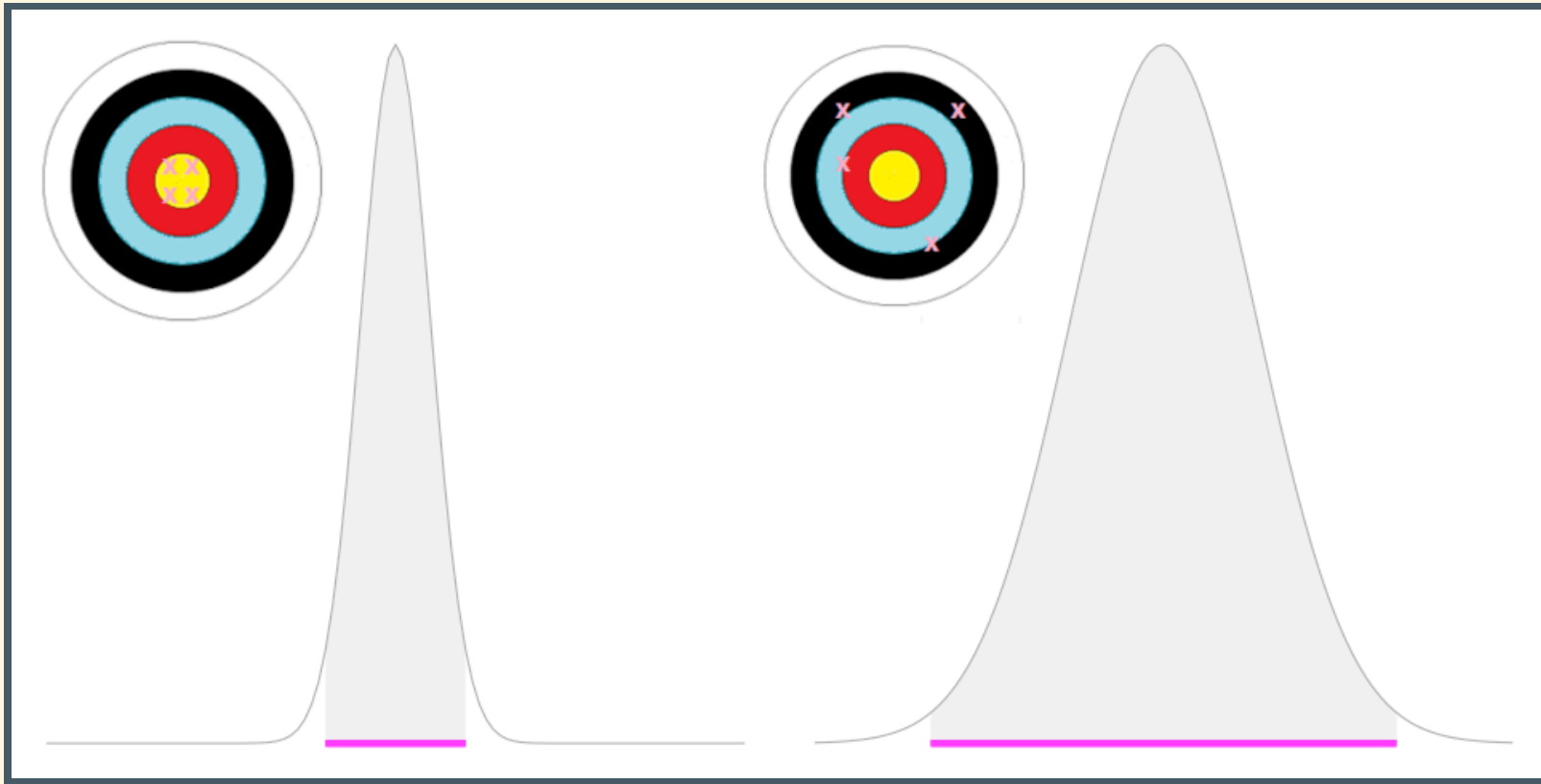
Esercizio #5 -- Soluzione

? Più l'intervallo di confidenza è largo

- a) meno siamo precisi ☒
- b) più siamo precisi
- c) non c'è differenza
- c) non posso rispondere



Esercizio #5 -- Soluzione



Esercizio #6

“ In media, quanti partner sessuali gli uomini inglesi, tra i 35 e 44 anni di età, riferiscono di aver avuto nella loro vita? ”

? Qual è il 95% CI?

$$n_{\text{uomini}} = 760$$

$$\bar{x} = 11.4$$

$$s = 11.2$$

$$SE = \sigma / \sqrt{n} = ?$$

Esercizio #6 -- Soluzione

“ In media, quanti partner sessuali gli uomini inglesi, tra i 35 e 44 anni di età, riferiscono di aver avuto nella loro vita? ”

? $n_{\text{uomini}} = 760$

$$\bar{x} = 11.4$$

$$s = 11.2$$

$$SE = \sigma / \sqrt{n} = ? \rightarrow \hat{SE} = s / \sqrt{n} = \frac{11.2}{\sqrt{760}} = 0.41$$

Esercizio #6 -- Soluzione

“ In media, quanti partner sessuali gli uomini inglesi, tra i 35 e 44 anni di età, riferiscono di aver avuto nella loro vita? ”

? $n_{\text{uomini}} = 760$

$$\bar{x} = 11.4$$

$$s = 11.2$$

$$SE = \sigma / \sqrt{n} = ? \rightarrow \hat{SE} = s / \sqrt{n} = \frac{11.2}{\sqrt{760}} = 0.41$$

$$\begin{aligned} 95\%CI &= (\bar{x} - 1.96 \times \hat{SE}; \bar{x} + 1.96 \times \hat{SE}) = \\ &= (11.4 - 1.96 \times 0.41; 11.4 + 1.96 \times 0.41) = \\ &= (10.6; 12.2) \end{aligned}$$

Esercizio #6 -- Soluzione

“ In media, quanti partner sessuali gli uomini inglesi, tra i 35 e 44 anni di età, riferiscono di aver avuto nella loro vita? ”

? $n_{\text{uomini}} = 760$
 $\bar{x} = 11.4$
 $s = 11.2$

$$SE = \sigma / \sqrt{n} = ? \rightarrow \hat{SE} = s / \sqrt{n} = \frac{11.2}{\sqrt{760}} = 0.41$$

$$95\%CI = (10.6; 12.2)$$

$$\text{via Bootstrapping} \rightarrow 95\%CI = (10.6; 12.1)$$

Esercizio #7

? Dato che $\mathcal{N} = (\mu, \frac{\sigma^2}{n})$ con $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \rightarrow$ standard error (SE),
come posso restringere l'intervallo di confidenza?

- a) aumentando n
- b) diminuendo n
- c) aumentando σ
- d) diminuendo σ
- e) nessuna delle precedenti
- f) non ho abbastanza elementi per rispondere

Esercizio #7 -- Soluzione

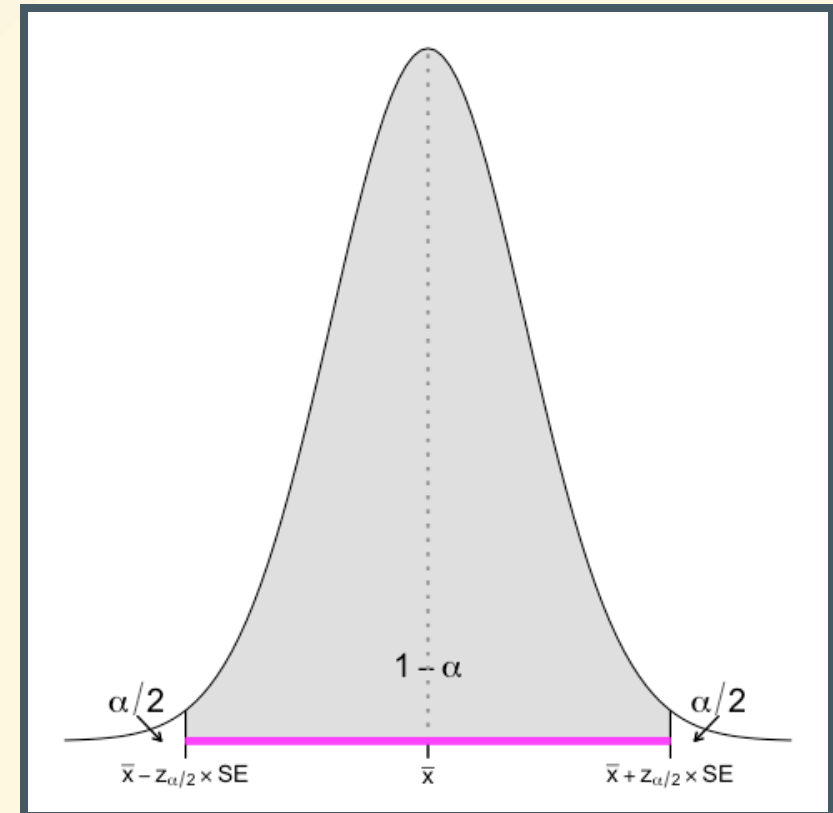
? Dato che $\mathcal{N} = (\mu, \frac{\sigma^2}{n})$ con $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \rightarrow$ standard error (SE),
come posso restringere l'intervallo di confidenza?

- a) aumentando n ☒
- b) diminuendo n
- c) aumentando σ
- d) diminuendo σ ☒
- e) nessuna delle precedenti
- f) non ho abbastanza elementi per rispondere

Il coefficiente di attendibilità α

🎯 95% CI = $(\bar{x} - 1.96 \times \hat{SE} ; \bar{x} + 1.96 \times \hat{SE})$
1.96 ?

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	

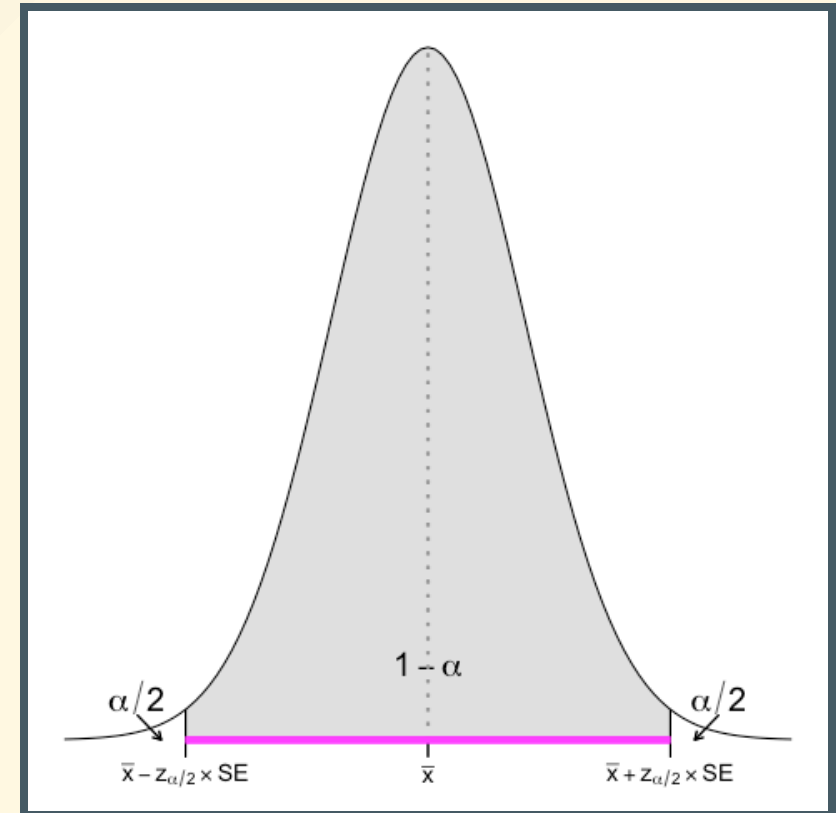


Il coefficiente di attendibilità α


🎯 95% CI = $(\bar{x} - 1.96 \times \hat{SE} ; \bar{x} + 1.96 \times \hat{SE})$
1.96 ?

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	

$$100\% - 2.5\% = 97.5\%$$



Il coefficiente di attendibilità α

 95% CI = $(\bar{x} - 1.96 \times \hat{SE} ; \bar{x} + 1.96 \times \hat{SE})$
1.96 ?

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	1.96

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9718	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

$$100\% - 2.5\% = 97.5\% \rightarrow z = 1.96$$

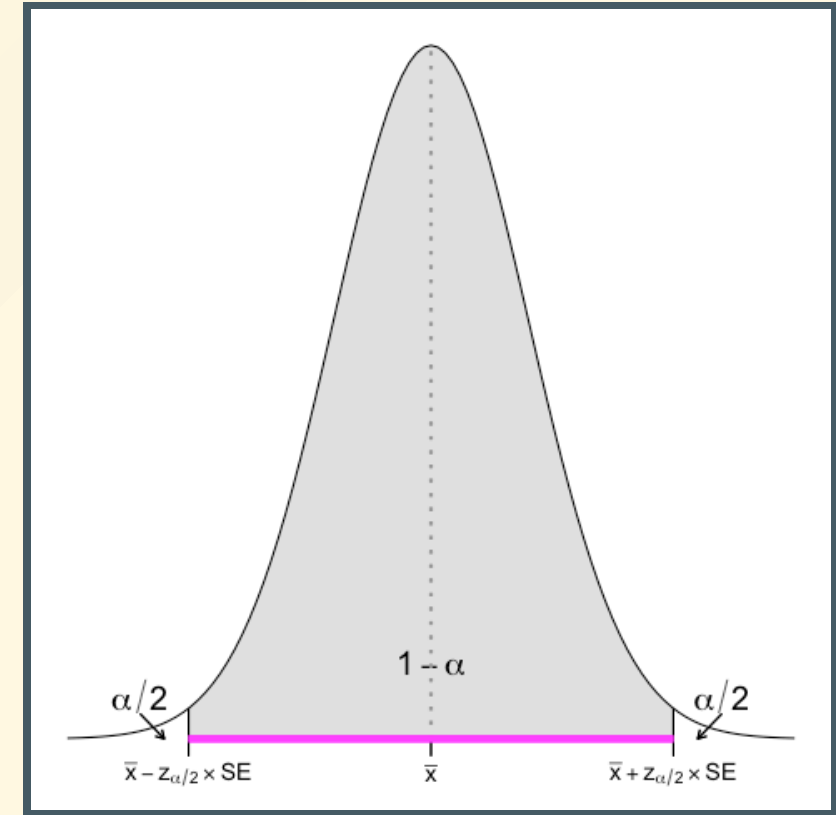
Il coefficiente di attendibilità α

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
95%	5%	2.5%	1.96
90%	10%	5.0%	1.65
99%	1%	0.5%	2.58

$$100\% - 2.5\% = 97.5\% \rightarrow z = 1.96$$

$$100\% - 5.0\% = 95.0\% \rightarrow z = 1.65$$

$$100\% - 0.5\% = 99.5\% \rightarrow z = 2.58$$



Una regola empirica

Il margine di errore (in percentuale) è al più $\pm 100/\sqrt{n}$

Esercizio #8

Il margine di errore (in percentuale) è al più $\pm 100/\sqrt{n}$

? Calcolate il 95% CI per le due affermazioni

05:00

**AZIONE ANTI-RUGHE
COMPROVATA**

Formula brevettata, specificatamente studiata per la delicata zona del contorno occhi:

- Riduce visibilmente le rughe intorno agli occhi del 17%**
- Pelle più liscia e tonica al tatto
- Riduce le occhiaie del 33%**
- Riduce il gonfiore sotto agli occhi

Subito	Il contorno occhi è nutrito intensamente
Dopo 2 sett.	Il contorno occhi è rivitalizzato e dall'aspetto più fresco.

*in-vitro **Studio di 2 settimane con 31 donne, 2019

Esercizio #8 -- Soluzione

Il margine di errore (in percentuale) è al più $\pm 100/\sqrt{n}$

? Calcolate il 95% CI per le due affermazioni

$$ME = 100/\sqrt{31} = 18\%$$

$$\begin{aligned} 95\% \text{ CI}_{\text{rughe}} &= (17 - 18; 17 + 18) \\ &= (-1\%, 35\%) \end{aligned}$$

$$\begin{aligned} 95\% \text{ CI}_{\text{occhi}} &= (33 - 18; 33 + 18) \\ &= (15\%, 51\%) \end{aligned}$$

**AZIONE ANTI-RUGHE
COMPROVATA**

Formula brevettata, specificatamente studiata per la delicata zona del contorno occhi:

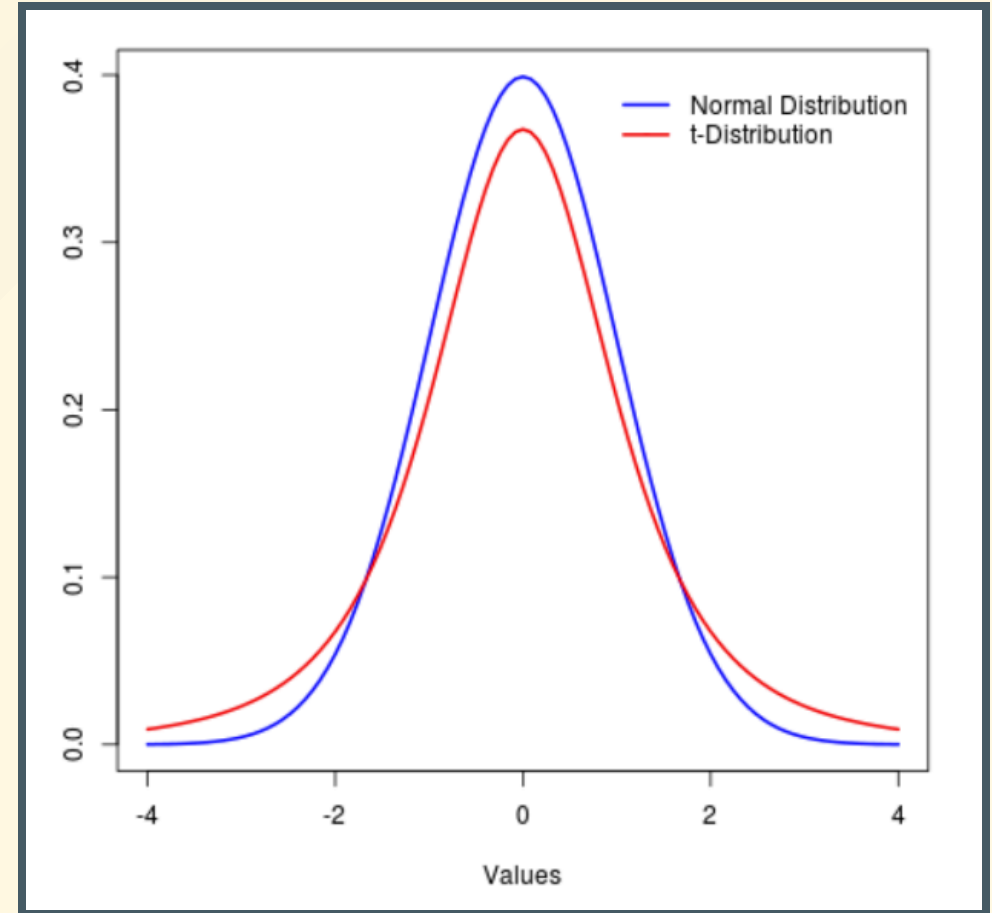
- Riduce visibilmente le rughe intorno agli occhi del 17%**
- Pelle più liscia e tonica al tatto
- Riduce le occhiaie del 33%**
- Riduce il gonfiore sotto agli occhi

Subito	Il contorno occhi è nutrito intensamente
Dopo 2 sett.	Il contorno occhi è rivitalizzato e dall'aspetto più fresco.

*in-vitro **Studio di 2 settimane con 31 donne, 2019

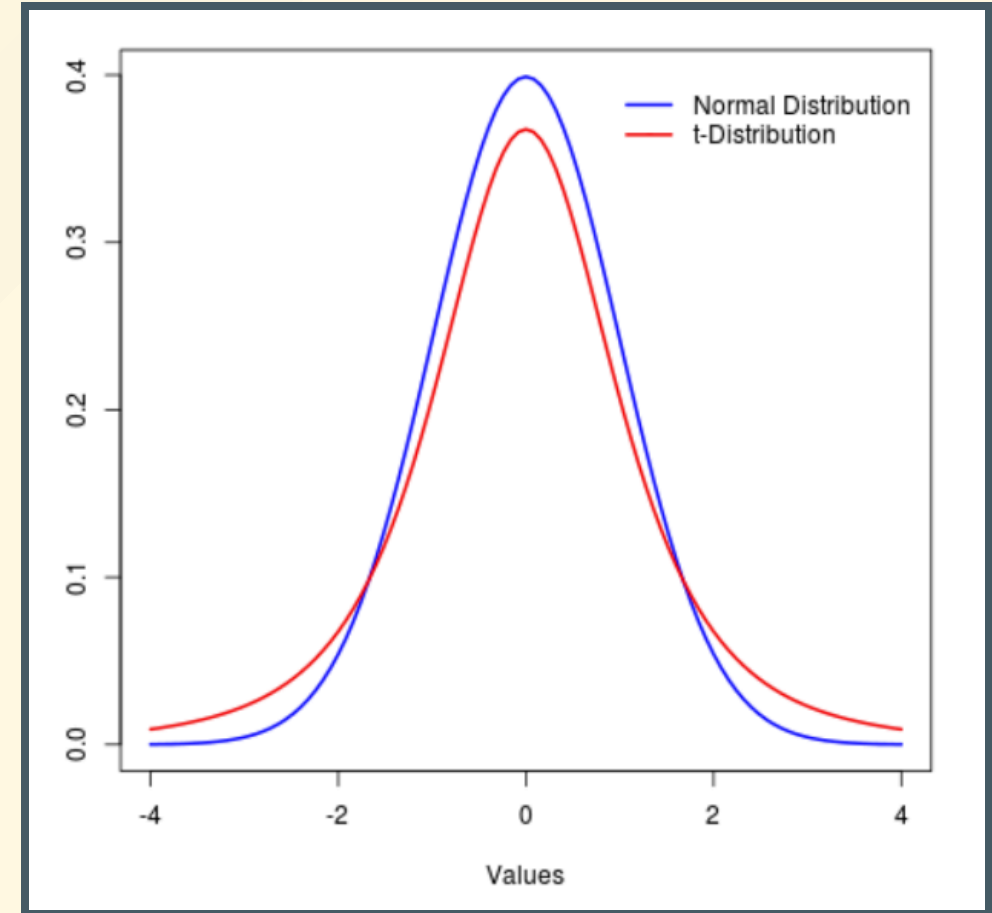
E se il campione è piccolo?

- Non posso approssimare a una normale
- Uso la t di Student



t di Student per campioni piccoli

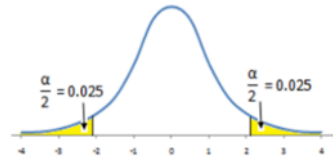
- Non posso approssimare a una normale
- Uso la t di Student
 - considera i gradi di libertà (df)
 - per un campione di dimensione $n \rightarrow df = n - 1$
 - per due campioni di dimensione $n_1 \wedge n_2 \rightarrow df = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$



t di Student per campioni piccoli

- Non posso approssimare a una normale
- Uso la t di Student
 - considera i gradi di libertà (df)
 - per un campione di dimensione $n \rightarrow df = n - 1$
 - per due campioni di dimensione $n_1 \wedge n_2 \rightarrow df = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$

$$95\% \text{ CI} = (\bar{x} - t \times \hat{SE}; \bar{x} + t \times \hat{SE})$$



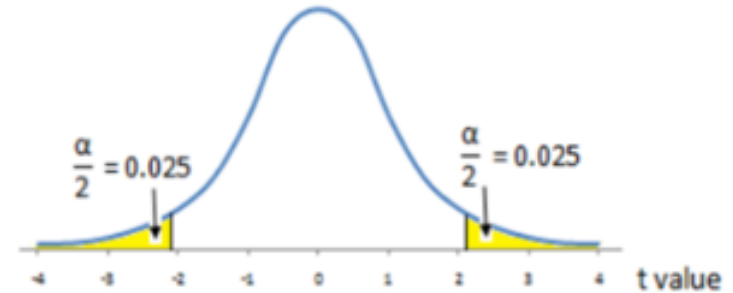
A graph of a t-distribution curve with the area under the curve in both tails shaded yellow. The x-axis is labeled 't value' and ranges from -4 to 4. The y-axis represents the probability density. The area in each tail is labeled $\frac{\alpha}{2} = 0.025$.

Confidence	90	95	95.45	99	99.73
alpha	0.1000	0.0500	0.0455	0.0100	0.0027
df					
1	6.314	12.706	13.968	63.657	235.784
2	2.920	4.303	4.527	9.925	19.206
3	2.353	3.182	3.307	5.841	9.219
4	2.132	2.776	2.869	4.604	6.620
5	2.015	2.571	2.649	4.032	5.507
6	1.943	2.447	2.517	3.707	4.904
7	1.895	2.365	2.429	3.499	4.530
8	1.860	2.306	2.366	3.355	4.277
9	1.833	2.262	2.320	3.250	4.094
10	1.812	2.228	2.284	3.169	3.957
11	1.796	2.201	2.255	3.106	3.850
12	1.782	2.179	2.231	3.055	3.764
13	1.771	2.160	2.212	3.012	3.694
14	1.761	2.145	2.195	2.977	3.636
15	1.753	2.131	2.181	2.947	3.586
16	1.746	2.120	2.169	2.921	3.544
17	1.740	2.110	2.158	2.898	3.507
18	1.734	2.101	2.149	2.878	3.475
19	1.729	2.093	2.140	2.861	3.447
20	1.725	2.086	2.133	2.845	3.422
21	1.721	2.080	2.126	2.831	3.400
22	1.717	2.074	2.120	2.819	3.380
23	1.714	2.069	2.115	2.807	3.361
24	1.711	2.064	2.110	2.797	3.345
25	1.708	2.060	2.105	2.787	3.330
26	1.706	2.056	2.101	2.779	3.316
27	1.703	2.052	2.097	2.771	3.303
28	1.701	2.048	2.093	2.763	3.291
29	1.699	2.045	2.090	2.756	3.280
30	1.697	2.042	2.087	2.750	3.270
40	1.684	2.021	2.064	2.704	3.199
50	1.676	2.009	2.051	2.678	3.157
60	1.671	2.000	2.043	2.660	3.130
70	1.667	1.994	2.036	2.648	3.111
80	1.664	1.990	2.032	2.639	3.096
90	1.662	1.987	2.028	2.632	3.085
100	1.660	1.984	2.025	2.626	3.077
1000	1.646	1.962	2.003	2.581	3.007
∞	1.645	1.960	2.000	2.576	3.000

t di Student per campioni piccoli

- Non posso approssimare a una normale
- Uso la t di Student
 - considera i gradi di libertà (df)
 - per un campione di dimensione $n \rightarrow df = n - 1$
 - per due campioni di dimensione $n_1 \wedge n_2 \rightarrow df = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$

$$95\% \text{ CI} = (\bar{x} - t \times \hat{SE}; \bar{x} + t \times \hat{SE})$$



Confidence	90	95	95.45	99	99.73
alpha	0.1000	0.0500	0.0455	0.0100	0.0027
df					
1	6.314	12.706	13.968	63.657	235.784
2	2.920	4.303	4.527	9.925	19.206
3	2.353	3.182	3.307	5.841	9.219
4	2.132	2.776	2.869	4.604	6.620
29	1.699	2.045	2.090	2.756	3.280
30	1.697	2.042	2.087	2.750	3.270
40	1.684	2.021	2.064	2.704	3.199
50	1.676	2.009	2.051	2.678	3.157
60	1.671	2.000	2.043	2.660	3.130
70	1.667	1.994	2.036	2.648	3.111
80	1.664	1.990	2.032	2.639	3.096
90	1.662	1.987	2.028	2.632	3.085
100	1.660	1.984	2.025	2.626	3.077
1000	1.646	1.962	2.003	2.581	3.007
∞	1.645	1.960	2.000	2.576	3.000

Cosa abbiamo imparato in questa lezione?

- Gli intervalli di confidenza (CI)/margin di errore sono un aspetto importante di come vengono comunicate le statistiche
- La dimensione del campione influenza la larghezza dei CI
- Attraverso il bootstrapping si ricampiona il dataset originale con rimpiazzo, ottenendo distribuzioni che tendono alla normale
- Il teorema del limite centrale ci dice che le distribuzioni campionarie tendono alla normale per campioni grandi, con formule per calcolare i CI
- Un CI del 95% risulta da una procedura che nel 95% dei casi contiene il valore della popolazione
- Per campioni piccoli, la distribuzione campionaria viene approssimata dalla distribuzione t di Student